14

Mediation Analysis with Count Outcomes

What counts can't always be counted; what can be counted doesn't always count

- ALBERT EINSTEIN

Navigation Tips

On the *Programs* tab of my website, programs are provided to conduct non-Mplus analyses in this chapter. Most programs have video links that briefly explain the method, show how to use the program, and review output. I provide links to the videos throughout this chapter. On the *Resources* tab of my webpage, scroll to

Chapter 14 for links to relevant online tutorials. To import data from other stat packages into R format for my programs, use the *Import/Syntax* tab on my website. To learn the basics of Mplus syntax, see *Chapter 11* and the *Syntax* tab on my website. For general Mplus output structure, see the *Output* tab on my website (hover the cursor over output text for tooltips).
*** To open a hyperlink in the same window, click on it. To open it in a new background tab, press Ctrl+click (for Macs, Cmd+click). To bring it to immediate focus in a new tab when you activate it, press Ctrl+shift+click (or Cmd+Shift+click for Macs) on the link. To search for a term within an open pdf or on a web page, use Ctrl+f (for Macs, Cmd+f). ***

INTRODUCTION

Researchers often study what are known as **count outcomes**. Count outcomes are bounded by zero and often are non-normally distributed with substantial skew. This can affect the accuracy of significance tests and confidence intervals if standard regression methods are applied to them. Count outcomes also are discrete rather than continuous in character. Statisticians have developed what they call **count regression models** to deal with the above complications. More generally, they are called **discrete outcome regression models**.

Consider a count outcome like the number of times young adolescents have used marijuana. The distribution of this variable has many observations at the value of zero (because most adolescents do not smoke marijuana), fewer observations at the value of one (some have tried marijuana once), still fewer observations at the value of two (some have smoked it a couple of times), and so on. Such skewed distributions are typical of many, but not all, count outcomes.

The use of the term "count outcome" is a bit of a misnomer because many of the variables we treat as continuous are technically counts. The number of dollars and cents that someone earns in a year technically is a count but we analyze it as if it were continuous. Nor does income necessarily have the kind of distribution typical of traditional "count" data. When the number of discriminations on a quantitative dimension becomes sufficiently large, it often is possible to analyze it using traditional methods that assume continuous outcomes. I discuss this matter in somewhat more depth in Chapters 3 and 13.

TYPES OF DISTRIBUTIONS THAT CHARACTERIZE COUNT OUTCOMES

Statisticians have identified a wide range of distribution types that often characterize counts, In this section, I consider two of the more popular distributions, the Poisson distribution and the negative binomial distribution.

The Poisson Distribution

One type of distribution that may be applicable to count data is the Poisson distribution. Like the normal distribution, there is not one Poisson distribution, but a family of Poisson distributions that all share common features. For example, the mean of a Poisson distribution always equals its variance.

The formal mathematical representation of Poisson distributions is provided in Long (1997) and I do not delve into its mathematical underpinnings here. Figure 14.1 shows examples of Poisson distributions with different means, ranging from 1 to 6. Note that as the mean of the distribution becomes larger than 6, the shape of the distribution looks more and more like a normal distribution. A Poisson distribution essentially becomes bell shaped and symmetric when it has a mean of 10 or greater.

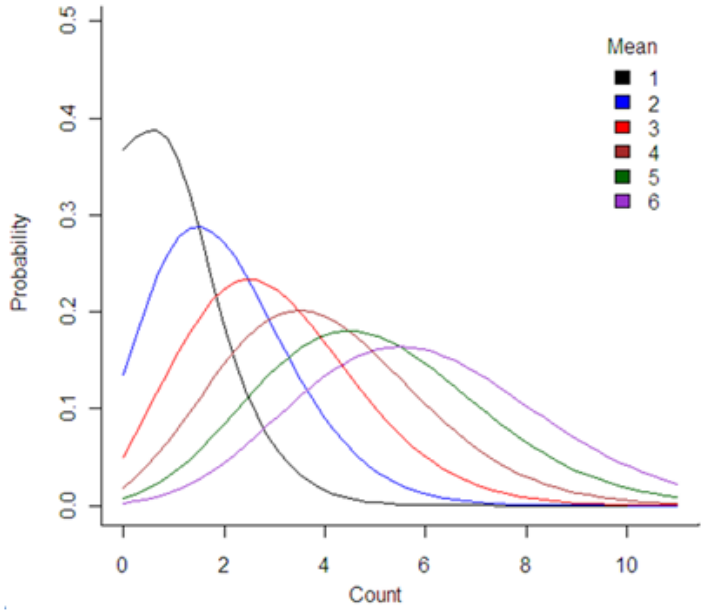


FIGURE 14.1. Example Poisson distributions

Suppose the mean count for 15 years old adolescents is 1. If the data for these

adolescents are Poisson distributed, the counts will follow the distributional form illustrated in Figure 1 for μ =1. Suppose the mean count for 16 years old adolescents is 2. It should follow the distributional form for μ = 2 in Figure 1. And so on.

Examine the Poisson distribution in Figure 1 where the mean is 1.0. Note there are many zeros and as one moves away from zero, the probability of higher valued counts becomes smaller. This is roughly the dynamic I described above for marijuana use in adolescents. So perhaps marijuana use is Poisson distributed. Certainly, this is a more reasonable assumption than assuming the distribution is normal.

If you are working with a count outcome you think is Poisson distributed, an informal way to determine if this is the case is to calculate its mean and variance. They should be equal. If the variance is larger than the mean, then the distribution is said to be **overdispersed**. If the variance is smaller than the mean, the distribution is said to be **underdispersed**. Of course, in sample data, the mean and variance may not be exactly equal because of sampling error.

In Poisson regression, we regress a count outcome Y onto a set of predictors, much like we do in traditional OLS regression. If the distribution of Y at each predictor profile is indeed Poisson distributed, then "Poisson regression" can be used to model the count means and to make statistical inferences from sample data to populations. Inferential tests involving the Poisson distribution are well developed.

Additional Considerations when Deciding to Use Poisson Regression

In addition to making the assumption that data are Poisson distributed, there is an additional consideration that should be taken into account when deciding to apply Poisson regression if one or more of the predictor variables are many valued and quantitative. Let's again use age as an example. If we calculate a mean count for 12 year old adolescents, a mean count for 13 year old adolescents, a mean count for 14 year old adolescents, and so on, Poisson regression makes an assumption about how the mean counts vary across values of age. In traditional OLS regression, the assumption is that the mean of Y is a linear function of the X scores. In Poisson regression, this is not the case. Let μ_i be the mean count for a given predictor profile *i*. Poisson regression assumes the various μ_i and X are related by a **log function**, specifically:

$$\log(\mu_i) = \alpha + \beta X_i$$

where log refers to the natural log and α and β are "adjustable constants" representing an intercept and slope that modify the logarithmic curve that characterizes the relationship between variations in X and the mean count.

As an example, suppose the mean counts for the number of times adolescents have smoked marijuana as a function of age are as follows (with logs of the mean counts shown to the right):

Age	Mean Count (μ_i)	Log(μ_i)
12	0.2019	-1.6
13	0.2466	-1.4
14	0.3012	-1.2
15	0.3679	-1.0
16	0.4493	-0.8
17	0.5488	-0.6

The mean counts are less than one in this example because most adolescents, no matter what their age, have not smoked marijuana and the data are dominated by zeros. Examine the column for the log of the μ. In this case, the log of the μ_i is a linear function of age with an intercept of -4.0 and a slope of 0.20. For every one unit age increases, the log of the mean count increases by 0.20 units:

$$\text{Log}(\mu_i) = \alpha + \beta X_i = -4.0 + 0.20 \text{ Age}$$

The point I emphasize here is that although there is a linear relationship between age and log(μ_i), the relationship between age and the mean count is non-linear.

If the relationship between X and log(μ_i) is *not* linear in the way implied by a Poisson regression model, then using Poisson regression is not appropriate because it is misspecified; the function surrounding the mean structure is misspecified. I might still use Poisson regression in the face of such misspecification if I believe the degree of departure from the assumed function is minimal or non-consequential. If it is consequential, then I need to deal with the non-linearity between X and log(μ_i) by, say, adding a quadratic X^2 term to the equation, or using some other strategy to correctly specify the model.

A problem I sometimes encounter with multiple continuous predictors in Poisson regression is that the function relating the mean counts to X is Poisson for some of the continuous predictors but not for others. One solution sometimes used is to mathematically transform predictors so that they are linearly related to the log Y. but doing so can undermine the interpretability of the coefficients associated with them.

It is possible in some scenarios that the Poisson model indeed describes the mean structure of Y across predictor profiles, but the distribution of the Y scores might not quite be Poisson distributed at different predictor profiles. Perhaps the Y scores are close to being Poisson distributed, but there is some overdispersion present. One strategy for dealing with such cases is to use robust estimators for standard errors so the significance tests and confidence intervals are not adversely affected by the overdispersion. This approach is referred to as **robust Poisson regression** and the robust standard errors used are typically Huber-White sandwich estimators.

Yet another strategy for over-dispersion is to use the Poisson model, but to modify it in a way that includes a parameter that adjusts for over- or under-dispersion. This approach is called **quasi-likelihood Poisson regression** and uses a model that is more flexible than the traditional Poisson model (which restricts means and variances to be equal). I do not delve into the mathematics of quasi-likelihood Poisson regression here (see Cameron and Trivedi, 1998, for details). However, it is mis specified another approach for dealing with dispersion issues in Poisson regression.

A final approach to dealing with dispersion issues in Poisson regression uses a **random intercept Poisson model** as proposed by Muthén, Muthén and Asparouhov (2016). Using an SEM framework, these authors introduce a continuous latent variable, *f*, into a cross-sectional model that implements a Poisson-lognormal (PLN) mixture model to account for unobserved heterogeneity and overdispersion. By adding *f*, Muthén et al. allow each individual to have their own unique "multiplier" on their expected count rate, which mimics a random intercept, thereby absorbing the excess variance (overdispersion) that the predictors cannot account for.

The *f* parameter has desirable properties in SEM frameworks in that one can specify predictors of *f* or use *f* to predict substantively interesting outcomes. In this sense, overdispersion becomes a meaningful substantive variable in a larger structural model. Also, if you have multiple count outcomes in a model, you can correlate their respective *f* variables to determine if their unexplained overdispersions are related. The Muthén et al. latent *f* approach has rarely been applied to count regression scenarios, but it is an interesting strategy for dealing with overdispersion and one that differs from the negative binomial approach I outline in the next section.

The Negative Binomial Distribution

An alternative distribution to the Poisson distribution is the negative binomial distribution. When invoked, we use what is called a **negative binomial regression model**. The **negative binomial distribution**, like the normal distribution, is a family of distributions that has many different forms. There are two key parameters that impact its form, a mean and **theta**, which also is called the **shape parameter**. In a negative binomial distribution, a mean does not necessarily equal its variance. For a technical description of theta and the mathematics of the negative binomial distribution, see Long (1997). Figure14.2 presents examples of negative binomial distributions with different means and theta = 1; Figure 14.3 presents examples with a mean of 2 and different theta values.

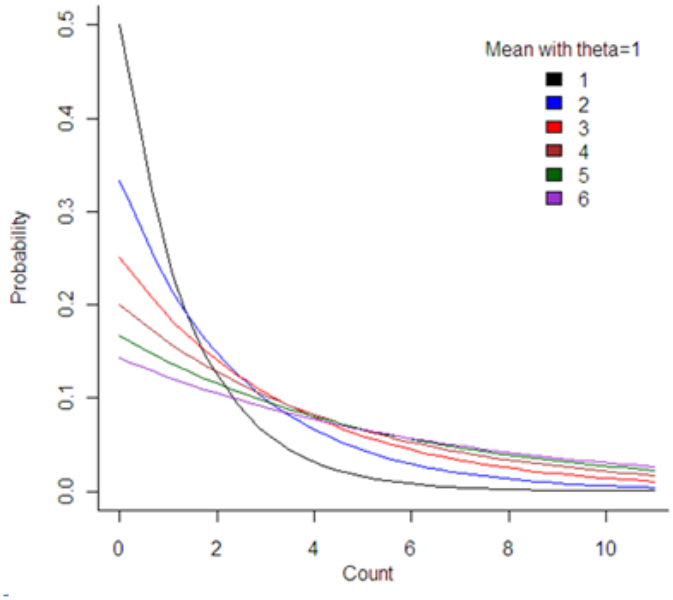


FIGURE 14.2. Negative binomial distributions with different means

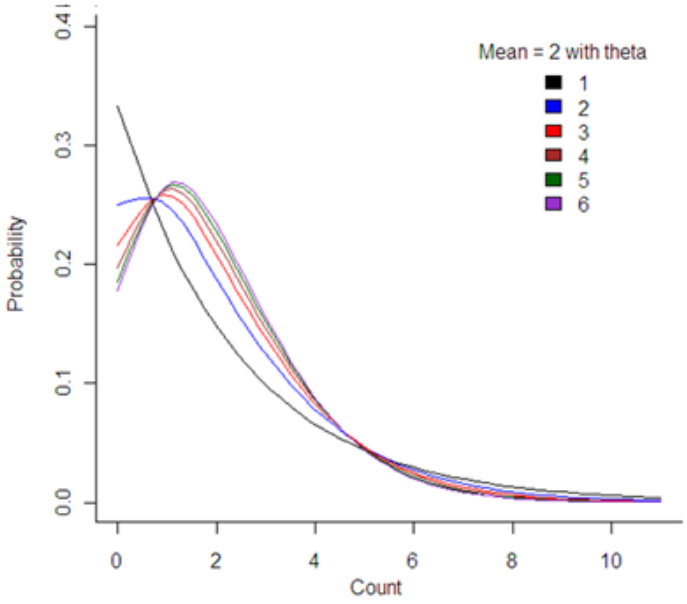


FIGURE 14.3 Negative binomial distributions with different thetas

When applying negative binomial regression, we assume the Y scores are distributed in accord with a negative binomial distribution at any given predictor profile. It turns out that

across the values of a continuous predictor, X, negative binomial regression, like Poisson regression, assumes that log(μ) is a linear function of X. So the choice between Poisson regression and negative binomial regression is largely dictated by the presumed distribution of the counts at each of the predictor profiles.

There are other possible distributions than the Poisson and negative binomial that are used in count regression, and I discuss these below. Before doing so, however, I explore one of the key diagnostics for choosing between different count regression types.

COMPARING OBSERVED FREQUENCES FOR EACH COUNT WITH THEORETICALLY EXPECTED FREQUENCIES FOR EACH COUNT

Suppose I collect a set of sample data and calculate a frequency distribution for the outcome variable. I might find that for 1,000 youth, 730 of them have never tried marijuana, 122 of them have tried marijuana once, 55 have tried marijuana twice, and so on. This is referred to as the **observed frequency distribution** for the count values. Using methods described by Long (1997), we also can calculate a predicted frequency for each count based on the regression model we fit to the data using the particular distribution type we decide to explore, such as the Poisson distribution. This is called the **predicted frequency distribution** for the counts based on the regression model. If the Poisson regression model is, in fact, appropriate, there should be close correspondence between the predicted and observed frequency distributions.

Figure 14.4 presents a plot of the observed and predicted frequency distribution for a Poisson regression model that predicts the number of articles professors publish from their marital status, the number of children they have, years since obtaining their doctorate, gender, and the productivity of their mentor. Note the marked under-prediction of the frequency of zero counts. A regression model that assumes Poisson distributed data does not seem appropriate. Contrast this plot with a similar plot for a negative binomial model in Figure 14.5. These results support the choice of a negative binomial model.

Just because there is close correspondence between the predicted and observed frequency distributions does ***not*** mean the predictors are strongly related to the outcome. It only means the presumed distribution of counts (in this case, a negative binomial distribution) is reasonable.

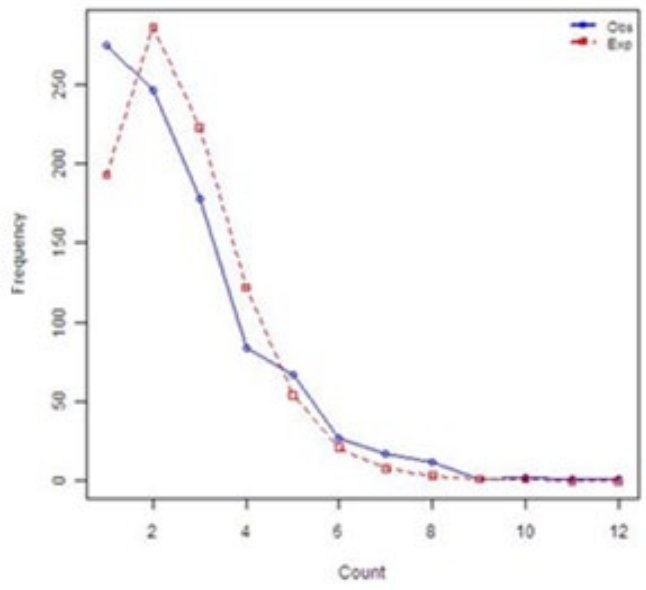


FIGURE 14.4 Predicted versus observed frequencies for a Poisson distribution

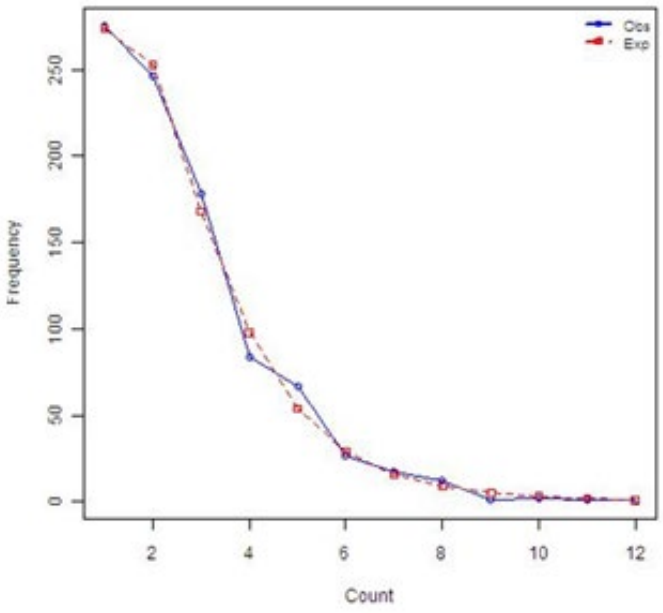


FIGURE 14.5 Predicted versus observed frequencies for a negative binomial distribution

ADDITIONAL PLAUSIBLE COUNT DISTRBUTIONS

My discussion thus far has focused on Poisson and negative binomial regression models. In this section, I discuss other distributions that can be used in count regression.

Zero-Inflated Models

A possible source of over-dispersion in Poisson models is an excess of zeros produced by some theoretical mechanism other than those that contribute to a Poisson process. When such a mechanism is present, we refer to this as a zero-inflated Poisson model and invoke such a distribution for analysis of the count data. In essence, we perform what is known as a **zero inflated Poisson regression**.

Zero-inflated models are essentially mixture models that combine two models into one integrated analysis. One model is called a **zero massing model** and this is combined with the more traditional Poisson model. In this modeling framework, there are two sources of zeros, (1) the Poisson process governing the counts, coupled with (2) the zero massing process. A binary regression model is used to represent the zero massing process (zero versus any positive count; this usually takes the form of a logit or probit model). A Poisson model is used for the count component.

When performing a zero inflated regression analysis, one literally specifies two predictor sets, one for the zero massing model and one for the count model. The predictor sets can be the same or different. The output of the analysis will be a binary regression that represents the zero massing process and a traditional Poisson regression that captures the count process.

To make the above concrete, consider a manufacturing example where the outcome variable is a count of the number of defects that products have after being manufactured on an automated assembly line. Some products have 0 defects, some have 1 defect, some have 2 defects, and so on. When all equipment in the manufacturing process is operating perfectly, it is impossible for a product to have a defect. However, if the equipment becomes misaligned, then a defect ***could*** occur (although it does not have to). Now, suppose the equipment is misaligned. In this case, the number of defects might follow a Poisson distribution. In this scenario, products with zero defects will still be produced, but the frequency with which defects occur in products will be dictated by a Poisson distribution, i.e., most products will have no defects, a smaller number of products will have one defect, an even smaller number of products will have two defects, and so on, perhaps like the curve with the highest peak in Figure 1. This represents one process contributing to the observed distribution of defects.

The second process contributing to the count distribution is whether the equipment is in perfect alignment or not. If the equipment happens to be in perfect alignment, then zero defects occur for all products and this results in the number of zeros to "mass" or increase over and above the Poisson process that occurs during misalignment. The zero massing or 0-

1 part of the zero-inflated model maps onto the dichotomous aligned/not aligned dynamic. The "count" part of the zero-inflated model maps onto what operates when the equipment is misaligned. Note that in this type of model, even if the equipment is misaligned, it is still possible for zero defects to occur. This is a key characteristic that separates zero inflated models from another two-component regression model that can be used, called a hurdle model, which I discuss later.

Here is another example of the dynamic. Suppose I seek to model heavy drinking on the part of adolescents and ask a question about the number of drinks adolescents have consumed in the past 30 days. I will undoubtedly observe a large number of zeros in the response distribution. There are two sources of these zeros. One source is youth who are committed non-drinkers and who simply do not drink alcohol at all. The other source is youth who are drinkers but who just happen not to have consumed any alcohol in the past 30 days. The former source represents the zero massing model and the second source represents the Poisson process.

One typically invokes a zero inflated Poisson model if (1) it makes theoretical sense to do so and (b) there is close correspondence between the predicted and observed frequency distributions of the counts when a zero inflated Poisson model is used. Some researchers are quick to adopt a zero inflated Poisson model without thinking through whether the presumed process is appropriate to the problem at hand.

Negative binomial regression also can be subjected to point massing phenomena, so there also is a zero inflated negative binomial regression model.

For statistical details on zero inflated models, see Long (1997), Long and Freese (2006), and Cameron and Trivedi (1998, 2005). The unique feature of these models is that you specify two regression equations instead of one, namely a regression equation for each component. They are then simultaneously fit to the data. The statistical output from a zero-inflated regression analysis is two regression equations each with its own predictors and coefficients, (a) a logistic/probit model for the zero massing dynamic, and (b) a traditional Poisson regression model for the count dynamic. Despite distinct output, the equations have been estimated taking into account the full multivariate patterning of the data.

Hurdle Models

Another class of models for dealing with excess zeros is **hurdle models** (Cameron & Trivedi 1998, 2005). Like the zero inflated models, these are two-component models, but the assumed theoretical mechanisms that underlie the inflation of zeros are different than traditional zero-inflated models. Hurdle models tend to be used in economics. I personally believe they have wider applicability than the more commonly used zero inflated models.

The idea behind a hurdle model is that the movement from 0 to 1 is an initial "hurdle" one must cross and the processes that govern the "jumping of such hurdles" are (possibly)

different than those governing the frequency of engaging in the behavior once a person has crossed the hurdle. Consider the case of adolescent sexual behavior. When adolescents are virgins, it is a big step or "hurdle" to have one's first act of intercourse. Scientists have identified numerous predictors of this. Once adolescents have crossed the hurdle of first intercourse, then the number of lifetime instances of intercourse throughout later adolescence can follow, say, a truncated negative binomial function, but it can never be zero. The frequency of sexual intercourse is, in essence, truncated above zero. The predictors of frequency of sex at this point might be different, might overlap, or might be the same as those for loss of one's virginity. In hurdle models, zeros come only from the first step.

As another example, consider the number of annual emergency room (ER) visits one makes. The hurdle is that of becoming sick or injured badly enough to go to the hospital. If you don't cross this threshold, your count is 0.Once you crossed the hurdle and enter the ER, you are officially admitted. You cannot have "zero" visits once you are there and in the future, you may have additional visits.

In hurdle models, zeros come from a single process. In zero-inflated models, they come from multiple processes. You use zero inflated models when your zeros come from two processes, namely people who can *never* experience the event (called **structural zeros**), and those who can but just didn't during the study (called **sampling zeros**). For hurdle models, zeros only come from structural zeros. A hurdle model is used when everyone belongs to the same general population, but there is a distinct physical, psychological, or institutional barrier (the "hurdle") they must cross before a positive count can begin accumulating. Once crossed, a zero count is impossible.

Zero-Truncated Models

In some situations, we work with data where the probability of a count of 0 is theoretically nil. For example, we might predict the number of days that patients stay in a hospital, where 1 day is the minimum value that can occur. In these cases, one might apply a zero truncated version of Poisson regression or a zero truncated variant of negative binomial regression. See Cameron and Trivedi (1998, 2005) for details of these models.

Robust Linear Regression

Most of the above count regression models assume a linear relationship between continuous predictors and the log of the mean count across predictor profiles defined by those predictors. As noted, this implies a non-linear relationship between X and the mean counts across X. But what if the relationship between X and the mean counts is linear rather than non-linear? In this case, traditional count models are misspecified. Sometimes the specification error can have minor consequences, other times it can be consequential.

In situations where it is consequential, some researchers apply traditional linear regression to count data, but use robust algorithms, such as bootstrapping or a Huber-White sandwich estimator. If counts are near zero, then these approaches can yield predicted counts that are negative, a property that is bothersome to some. However, if the focus is on estimating coefficients in a causal effect framework rather than in a purely predictive capacity, the unbounded nature of the predicted values in can be less problematic (see Angrist & Pischke, 2008).

RESIDUAL TERMS AND SQUARED CORRELATIONS IN COUNT MODELS

For count outcomes, most statistical software does not report a residual term nor a squared multiple correlation. Count models, technically, have no random residual term because the defining equations are based on the distribution of the outcome variable itself (see Hilbe, 2011). Count regression models the mean directly using a link function and handles random variation via discrete probability distributions. Given this, there is no pure definition of unexplained variance or squared R for count models, although ad hoc approaches for pseudo-squared Rs have been proposed. For example, one can fit a count model with predictors and note the log likelihood that results and then refit the model but constrain the coefficients for the predictors to be zero so that an intercept only model is operative. The commonly reported McFadden pseudo squared R is defined as

$$1 - (LL_{FULL}/LL_{INTERCEPT})$$

where LL_{FULL} is the loglikelihood value for the model with covariates and $LLI_{INERCEPT}$ is the loglikelihood value for the intercept-only model.

You will encounter references to residuals in count regression but these refer to the disparities between predicted and observed mean counts. They do not concern the calculation of error variances. Such residuals are not needed to compute estimates for disturbance variance because there is no disturbance term in the model to begin with.

CHOOSING A COUNT REGRESSION MODEL

Which count model should one use? There is no simple answer to this question. Good analytic practice explores different plausible models. One diagnostic is to compare the observed and predicted frequency distributions of the counts, as discussed above. Another strategy sometimes used is to fit multiple models and then to compare them on their AIC and/or BIC indices. One then chooses the model with the best fit index. I return to this matter in the context of the numerical example below.

MODELS WITH OFFSETS: THE ANALYSIS OF RATES

Count data sometimes have an **exposure variable** associated with them that varies by individuals. An exposure variable, also called an **offset**, reflects the number of times the event could have happened for a given individual. For example, the count outcome might be the number of instances of unprotected sex an adolescent engages in during the past year and the offset might be the number of times the individual engaged in sex during the past year. One might be interested in examining the number of instances of unprotected sexual intercourse relative to the number of instances of sexual intercourse. To do so, we include the offset in the analysis, but in a special way (discussed shortly). When we conduct analyses with offsets, we are essentially analyzing a rate. In our example, the rate is the number of instances the individual has unprotected sex relative to the total number of instances of intercourse in which the individual engages.

When an offset is included in a Poisson or negative binomial regression model, the traditional practice is to set its regression coefficient to 1 and to log transform it. Why? A rate for the outcome variable, Y, at a given exposure value is defined as Count/Exposure. If we hold the exposure value constant at some value, E, then the average rate at predictor profile *i* is

$$\log(\mu_i / E) = \alpha + \beta X_i \qquad [14.1]$$

where μ_i is the mean count for predictor profile *i*. There are logarithm rules, one of which is known as the **quotient rule**. It states that

$$\log (Y/X) = \log(Y) - \log(X)$$

Applying this to the left hand side of Equation 14.1, yields

$$\log(\mu_i) - \log(E) = \alpha + \beta X$$

and if I then add log(E) to both sides of the equation, I get

$$\log(\mu_i) = \alpha + \beta X + \log(E)$$

which is the classic Poisson model but with a term added to the right hand side of the equation, log(E). This term is the log of the offset/exposure variable and it has an "implicit regression coefficient" of 1.0 because there is no coefficient to be estimated for it; it stands on its own. So, to model the rate, the offset/exposure variable needs to be log transformed and the coefficient for the offset needs to be constrained to equal 1. See Cameron and Trivedi (1998, 2005) for details about modeling with offsets.

Offset models are not typically applied to the zero inflated versions of the Poisson and

negative binomial models because of statistical complications that result from the two component structure of such models (see Cameron and Trivedi, 2005, for details). Nor can they be applied to hurdle models. Care also must be taken when applying offsets to negative binomial models even if they are not zero inflated. In negative binomial models, not only is the mean of the distribution affected by the offset, but theta also could be affected by it. This is problematic because theta effects are not part of a traditional model with an offset. This also is true for quasi-likelihood Poisson regression. The bottom line is the analysis of rates using offsets in the form described is really most appropriate with Poisson regression. In my experience, Poisson models rarely are applicable in practice, so this limits the offset strategy in count regression, accordingly.

An alternative approach to an exposure variable is to use it as a predictor in the model but without constraining its regression coefficient to be 1.0 but instead estimating it. In this case, one is no longer evaluating rates in the traditional count regression sense. One instead views the exposure variable as a potential determinant of the count. For example, if one is studying surgical complications at hospitals and the exponent of the coefficient of the exposure variable is greater than 1.0, then this means that hospitals that do more surgeries have a higher rate of complications per surgery. If the exponent of the coefficient is less than 1.0, then hospitals that do more surgeries have fewer complications per surgery.

In cases where the outcome variable takes the form of a proportion for each individual (e.g., the percent adherence exhibited by each individual to a treatment protocol but expressed in proportion form), an effective method of analysis is **fractional logit/probit modeling** (FRM, Papke & Wooldridge, 1996). This method uses a specialized form of a generalized linear model but with quasi-likelihood estimation. Instead of assuming a linear function between continuous predictors and the proportion constituting the outcome, FRM assumes an S-curve via a logit or probit link. It ensures that no matter how extreme an individual's data is, their predicted rate will never fall below 0 or exceed 1.0. If you believe the function linking your predictors to your outcome is S shaped, consider using this method. If you think it is linear, you might still be able to use it as long as the S is not pronounced or you can probably use a robust version of OLS regression. For the latter, know that predicted values outside the 0 to 1.0 range can occur; if too many occur, it can be problematic (see my discussion of this topic for the modified linear probability model in Chapter 12).

For both FRM and OLS, it sometimes is good practice to include the number of exposures/opportunities either as a control covariate or as a sampling weight (or both). Using it as a weight accounts for varying precision in the rate values given varying exposures/opportunities. For example, if one person has a rate of 0.40 based on five exposures (2/5) and another has the same rate based on 500 exposures (200/500), then the rate for the latter individual likely has less sampling error associated with it and this should be taken into account. In many cases, the number of exposures/opportunities is constant or near constant,

such as the adherence rate to a treatment protocol that has the same number of requirements for everyone (e.g., 10 practices they must do). In other cases, this will not be the case, such as the rate of unprotected sex for someone who engages in sex 3 or 4 times a month versus someone who engages in sex 25 or 30 times a month. Note that for scenarios where the number of exposures do not vary by much, the use of a robust estimator usually handles matters without the need for weights. But if the coefficient of variation for the number of exposures (the standard deviation of the exposures divided by the mean of the exposures) exceeds 0.10 or 0.15, you might need to introduce weights. Do not normalize or log transform the weights; use the sheer number of exposures when weighting (Papke & Wooldridge, 1996).

If you think the number of opportunities could influence your outcome, then you likely will want to include it as a predictor to adjust for possible confounding. When doing so, some researchers use a log transform for it, but others do not. The decision to do so depends on your underlying theory about exposure influence. If you believe that each unit increase in exposure adds a fixed, constant amount to the log-odds of the outcome, then you would not use a log transform. If, however, you believe the dynamic is non-linear in accord with a log function, you would use the transform. For example, if you expect a non-linear relationship where the difference between an exposure of 10 vs. 20 matters considerably but the difference between 500 and 520 matters little, one might use a log transform. Cases can arise where you use the number of exposures both as a weight and as a confound control.

I provide a program for fractional logit/probit modeling on my website called *Fractional logit regression* as well as a program for GAMs called *Generalized additive models*. I provide a worked example in the document *Fractional Logit Regression for Rates* on the Resources tab of my webpage for Chapter 14 and elaborate on many of the above issues as well as tests you use to explore specification error.

INTERPRETING COEFFICIENTS IN COUNT MODELS

In Poisson and negative binomial modeling, the coefficients that are reported are estimates of the coefficients in the equation

$$\text{Log}(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

The intercept is the predicted log of the mean count when all predictors equal zero. The β for a given X reflects the predicted change in the log of the mean count for every one unit that X increases, holding constant the other X in the equation.

If X is a dummy variable with 0,1 coding, then the coefficient is the log of the mean count for the group scored 1 minus the log of the mean count for the reference group.

Applied researchers sometimes find it difficult to interpret log mean differences, so it is not uncommon to "remove the logs" by calculating exponents (or anti-logs) of the intercept

and of the estimated β. When we calculate such exponents, the coefficients are interpreted using the logic of "multiplicative factors" rather than differences. Consider the estimated coefficients for the negative binomial model that predicted the number of publications of professors. Gender was coded 1 for females and 0 for males and the exponent of its coefficient was 0.82. This means the estimated mean number of publications by women was 0.82 that of men (i.e., the female mean is the estimated number for males multiplied by a factor of 0.82).

For the predictor of the number of publications of their mentor, the exponent of the coefficient was 1.03. This means that for every additional publication the mentor had, the estimated mean number of publications produced by the target professors increased by a multiplicative factor of 1.03. For example, if the estimated mean number of publications was 4.0 for professors with mentors who published 12 articles, than for those who had mentors with 13 articles, the mean number of publications is predicted to be (4.0)(1.03) = 4.12, and for those with mentors with 14 articles, it is (4.12)(1.03) = 4.24. For every one unit that X increases, the outcome mean changes by a multiplicative factor of the exponent of the coefficient associated with X.

When one conducts a Poisson or negative binomial count regression, the focus generally is on the exponents of the various coefficients in the model. Note the exponent of a coefficient of 1.00 is equivalent to a null effect – for every unit X increases, the mean count changes by a multiplicative constant of 1.00, i.e., it does not change. If the confidence interval for the exponent of a coefficient does not contain 1.0, the coefficient is statistically significant.

Coefficients in Models with Offsets

Coefficients in models with offsets have a special interpretation due to the presence of the offset in the predictor set. Traditionally, the intercept is the predicted log of the mean count when all predictors are zero. In models with offsets, the intercept is the predicted log of the mean count when all predictors are zero *and* the offset equals zero. If the offset was transformed to be log(offset), then a value of zero on log(offset) corresponds to a value of 1 on the non-transformed offset, because the log of 1 is zero. So for offset analysis, the intercept is the predicted log of the mean count when all the X predictors = 0 and the offset = 1.

The exponent of the coefficients in offset models is interpreted as described above, but now they apply to rates. For example, if the exponent of a coefficient for gender (where females are scored 1 and males are scored 0) is 2.0, this means the estimated rate was twice the size for females than it was for males, holding all other predictors constant. If the coefficient exponent for age is 1.5, then for every one year that age increases, the rate is predicted to increase by a multiplicative factor of 1.5. For a discussion of coefficient interpretation in fractional logit models, see the aforementioned document

AN RET EXAMPLE

I illustrate the application of SEM to an RET using hypothetical data for a an intervention focused on reducing binge drinking in college students. The outcome variable was the number of times students reported engaging in binge drinking in the past 30 days. The targeted mediators were the motivation to change their alcohol consumption measured on a -5 to +5 multi-item metric in which item responses were averaged (sample item: "How important is it for you to make any changes in your alcohol consumption?") and their self-perceived efficacy with respect being able to change their behavior if desired, also measured on a -3 to +3 metric (sample item: "How confident are you that you are capable of making any changes in your personal alcohol consumption?"). Measures of the mediators, which I refer to as MA and MB, were obtained at the immediate posttest and then the count outcome, Y, for self-reported binge drinking over the past 30 days was obtained 30 days later. Figure 14.6 presents the influence diagram for the model omitting correlations between exogenous variables but that nevertheless are modeled in my Mplus program. The disturbances for the mediators are allowed to correlate because of the likely existence of unmeasured shared influences on them.

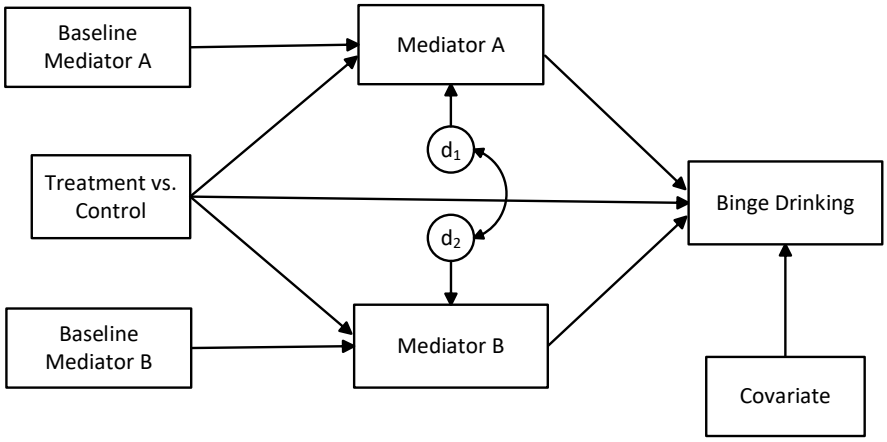


FIGURE 14.6 RET example with count outcome

Table 14.1 presents the Mplus syntax for purposes of model testing and analysis.

Table 14.1. Mplus Syntax for RET Example

```
1.  TITLE: Analysis of RET with Count Outcome ;
2.  DATA: FILE IS CountM.dat ;
3.  VARIABLE:
4.    NAMES = id y treat ma mb base_ma base_mb cov;
```

```
5.    USEVARIABLES =  y treat ma mb base_ma base_mb cov;
6.    COUNT = y(nb);
7.  ANALYSIS:  ESTIMATOR = MLR;
8.  MODEL:
9.    !Specify mediator regression equations
10.   ma ON treat base_ma (a1 c_ma);
11.   mb ON treat base_mb (a2 c_mb);
12.   !Specify outcome regression equation and label coefficients
13.   y ON ma mb treat cov (p_ma p_mb p_treat p_cov);
14.   !Specify and label intercepts of the mediators and outcome
15.   [y] (b0);  [ma] (b0_ma);  [mb] (b0_mb);
16.   !Specify and label sample means of baseline covariates
17.   [base_ma] (m_bma); [base_mb] (m_bmb); [cov] (m_cov);
18.   !Correlate the disturbances of ma and mb
19.   ma with mb ;
20. MODEL INDIRECT:
21.   y IND treat;
22. MODEL CONSTRAINT:
23.   NEW(irr_ma irr_mb irr_tr E_ma_c E_mb_c log_ctrl mean_ctrl
24.       E_ma_i E_mb_i log_int mean_int tot_log_d tot_irr tot_dif);
25.   !INCIDENCE RATE RATIOS FOR INDIVIDUAL PATHS
26.   irr_ma=EXP(p_ma); irr_mb=EXP(p_mb); irr_tr=EXP(p_treat);
27.   !CONTROL GROUP EXPECTED (MEAN) VALUES (treat = 0)
28.   E_ma_c = b0_ma+(c_ma*m_bma);
29.   E_mb_c = b0_mb+(c_mb*m_bmb);
30.   !Log mean count for control group
31.   log_ctrl = b0+(p_ma*E_ma_c)+(p_mb*E_mb_c)+p_treat*0+(p_cov*m_cov);
32.   !Estimated mean count for control group
33.   mean_ctrl = EXP(log_ctrl);
34.   !INTERVENTION GROUP EXPECTED MEAN VALUES (treat = 1)
35.   E_ma_i = b0_ma+a1+(c_ma*m_bma);
36.   E_mb_i = b0_mb+a2+(c_mb*m_bmb);
37.   !Log mean count for intervention group (includes direct path p_treat)
38.   log_int = b0+(p_ma*E_ma_i)+(p_mb*E_mb_i)+p_treat*1+(p_cov*m_cov);
39.   !Estimated mean count for intervention group
40.   mean_int = EXP(log_int);
41.   !TOTAL EFFECT
42.   !Total effect on the log scale
43.   tot_log_d = log_int-log_ctrl;
44.   !Total effect expressed as an overall ratio of the means
45.   tot_irr = EXP(tot_log_d);
46.   tot_dif = mean_int-mean_ctrl ; !Total effect expressed as a difference
47. OUTPUT: SAMP STDYX CINTERVAL RESIDUAL TECH4 ;
```

Most of the syntax should be familiar. Line 6 identifies the endogenous variable y as a count variable that is to be fit using a negative binomial (nb) model. For a Poisson model use (p), for a zero-inflated Poisson model use (i), and for a zero-inflated negative binomial model use (nbi). For hurdle models with Poisson counts use (ph) and for hurdle models with negative

binomial counts use (nbh). For zero inflated or hurdle models, such as y (nbi), the letter y when used in the MODEL command refers to the log rate of the count and y#1 refers to the probability of always being in the zero class. This notation also is used for hurdle models. Lines 25 and 26 in the MODEL CONSTRAINT command calculate the exponents of the three path coefficients for the prediction of Y from MA. MB and TREAT. These represent the relevant multiplying factors and sometime are referred to in the literature as **incident rate ratios**. Lines 27 to 29 estimate the control group means for MA and MB that are then used in Line 31. Line 31 calculates the log mean count for the control group by multiplying the path coefficients for the MA and MB predicting Y by the respective control group MA and MB means for each of them. Line 33 calculates the exponent of the Line 31 result to yield the predicted mean count for the control group. Lines 34 to 40 perform the same computations for the intervention group. Lines 45 calculates the total effect of the intervention on Y expressed as ratio of the intervention group mean count divided by the control group mean count and Line 46 calculates the difference between these two means rather than their ratio.

Model Fit

Mplus provides limited information about global model fit for count regression models, usually limited to log likelihoods and AICs and BICs. Consequently, you need to use somewhat ad hoc approaches to gain perspectives on fit. The tested model in Figure 14.6 is structurally just-identified with minor disparities from sample values occurring for the exogenous variables due to the use of numerical integration. If the structural model was not just-identified, I would examine tests of independence as discussed in Chapter 8, but this is not necessary here. Given this, I turn my attention to ensuring that the assumed negative binomial distribution is appropriate for modeling the data.

Mplus does not readily provide predicted versus observed counts for models of count outcomes with predictors so I often handle distribution examination using LISEM. Specifically, I use count regression programs in R focused on just the count portion of the model, in this case, the regression of binge drinking (Y) onto motivation to change (MA), self-efficacy (MB), and the treatment condition (TREAT). I provide a program on my website called *Count regression* on the Programs tab that accomplishes the analyses. Figure 14.7 presents the plot of predicted versus observed counts produced by the program for the negative binomial analysis. The fit of the distribution seems reasonable. I compared this model to an alternative distribution, namely a Poisson model, which I show in Figure 14.8. It is evident that the negative binomial model provides the better fit.

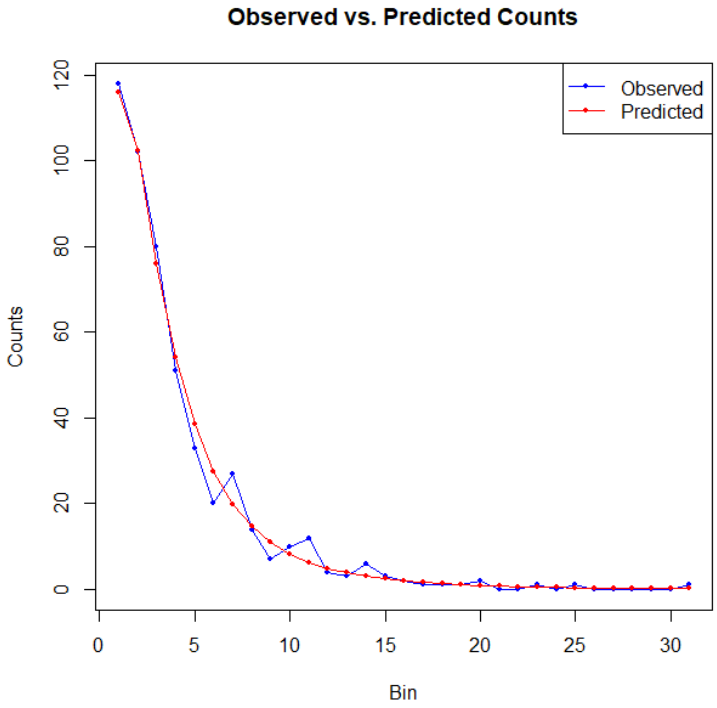


FIGURE 14.7 Predicted versus Obtained Counts for Negative Binomial Model

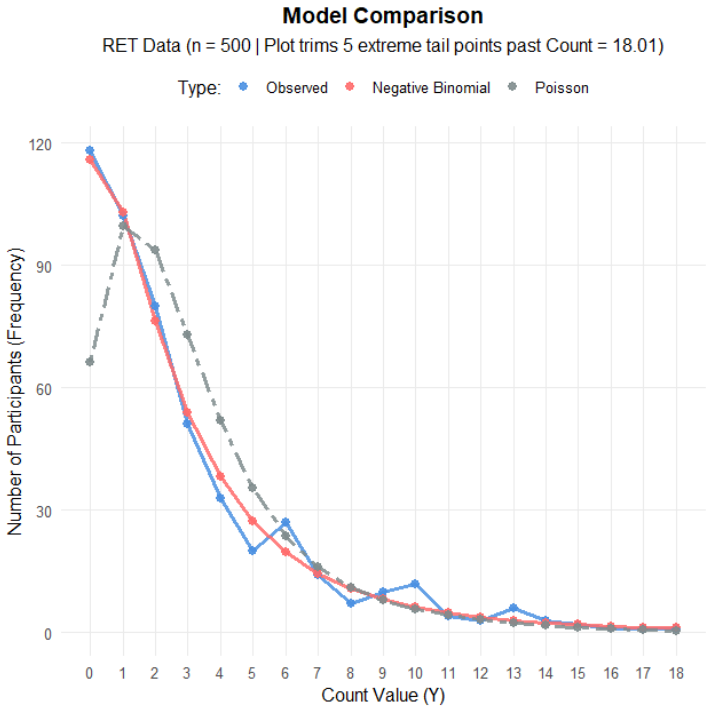


FIGURE 14.8 Comparison of Negative Binomial Model with Poisson Model

Some researchers compare different model fits using the respective BIC statistics for the models. In the current case, the BIC for the Poisson model was 2462.74 and for the negative binomial model it was 2137.53. The model with the lowest BIC yields the better fit. In this case, the disparity is quite large, favoring the negative binomial model (see Chapter 7). I explored other possible count distributions as well but ultimately concluded the negative binomial model was the most reasonable choice.

Effect of the Intervention on the Outcome

The estimate of the overall effect of the intervention on binge drinking is obtained from the `TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS` section. Here is the (edited) output:

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

Effects from TREAT to Y

  Total              -0.631      0.096     -6.552      0.000
```

The value -0.63 ±0.19 is the estimated log mean difference in binge drinking between the intervention and control groups (CR = 6.55, $p < 0.05$). The exponent of this value is the incidence rate ratio (IRR) and equaled 0.53 ± 0.10. This means that the mean count for binge drinking was cut in about half relative to the mean count for the control group.

I broke this result down more in the `MODEL CONSTRAINT` commands which yielded the following output:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
New/Additional Parameters

    MEAN_CTR           3.410      0.238     14.302      0.000
    MEAN_INT           1.815      0.136     13.342      0.000
    TOT_IRR            0.532      0.051     10.390      0.000
    TOT_DIF           -1.595      0.260     -6.131      0.000
```

The mean number of times the intervention group engaged in binge drinking in the past 30 days was 1.81 ±0.27. The corresponding mean for the control condition was 3.41 ±0.48. If I divide 1,815 by 3.410 I obtain 0.532, which is the incidence ratio. I also calculated the simple mean difference and it equaled -1.59 ±0.52. Interestingly, when discussing meaningfulness

standards with program staff and designers, they felt that cutting binge drinking nearly in half would be quite meaningful. But when told the mean number of binge drinks would be reduced on average by 1.5 days over the course of a 30 day period, they were not impressed. In other words, the way the effect size was framed impacted their evaluation of it. If we work from the perspective of incidence rate ratios, suppose the meaningfulness standard was to reduce binge drinking by one fifth or more (20% or 0.20). The confidence interval for the IRR was 0.43 to 0.63. Because the lower limit of the interval exceeded the standard, we declare the effect meaningful (to the chagrin, however, of those who thought in terms of absolute differences).

Technically, the confidence intervals for the IRR can be asymmetric which is best captured using bootstrapping. However, I often find the amount of asymmetry is relatively small and I use my judgment as to whether the more fine-grained reporting of them is substantively important. In the current case the above reported confidence interval was 0.43 to 0.63 whereas the bootstrapped intervals were 0.44 to 0.64, a difference that is not consequential. To conduct bootstrapping change `CINTERVAL` to `CINTERVAL(BOOTSTRAP)` on the `OUTPUT` line and change the `ANALYSIS` line to read `ANALYSIS: ESTIMATOR=ML; BOOT=1000 ;.`

Effect of the Intervention on the Mediators

Here are the results for the effect of the intervention on the two mediators, MA and MB:

```
MODEL RESULTS

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

 MA         ON
    TREAT             0.531      0.072      7.392      0.000
    BASE_MA           0.528      0.034     15.746      0.000

 MB         ON
    TREAT             0.498      0.076      6.544      0.000
    BASE_MB           0.461      0.038     12.092      0.000
```

The mean difference between the intervention and control group for MA on its -5 to +5 metric (with a pooled within-group covaried adjusted standard deviation of 0.80)[1], was 0.53 ±0.14 (CR = 7.39, $p < 0.05$). Suppose prior to the study the program staff decided that a meaningful mean difference is a third of a scale metric or 0.33. The confidence interval for the

[1] I obtained this value as the square root of the residual variance for MA as reported on the Mplus output in the `RESIDUAL VARIANCES` section of the `MODEL RESULTS`.

effect of the treatment on MA was 0.49 to 0.67. Because the lower limit of the interval exceeds that of the meaningfulness standard, the effect is declared meaningful.

I structed the MODEL CONSTRAINT commands to provide the mean MA values for the intervention and control conditions. Here are the results:

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

New/Additional Parameters
    E_MA_C            -0.262      0.057     -4.623      0.000
    E_MA_I             0.269      0.056      4.843      0.000
```

The estimated mean MA for the control group was -0.26 ±0.11 and for the intervention group it was 0.27 ±0.11.

Because I raised the issue of effect size interpretation earlier, it also may be instructive to calculate some of the more traditional effect size indices used in the social sciences for the above mean difference. Cohen's *d* is the estimated covariate adjusted mean difference divided by the within-group covariate adjusted standard deviation = 0.531/0.804 = 0.67. Many researchers would interpret this value as a medium effect size. The probability of exceptions to the rule is 0.32 (which I calculated using the probability of exceptions programs on the Programs tab of my webpage). This means that if you randomly select a person from the intervention group and a person from the control group, the person from the control group will have a *higher* MA score about 32 percent of the time with successive draws. Such cases represent exceptions to the rule that "people in the intervention group score higher on MA than people in the control group."

The mean difference between the intervention and control group for MB on its -5 to +5 metric (with a pooled within-group covaried adjusted standard deviation of 0.85), was 0.50 ±0.15 (CR = 6.54, $p < 0.05$). Suppose prior to the study the program staff decided that a meaningful mean difference is a third of a scale metric or 0.33. The confidence interval for the effect of the treatment on MA was 0.45 to 0.65. Because the lower limit of the interval exceeds that of the meaningfulness standard, the effect is declared meaningful.

As with MA, I structed the MODEL CONSTRAINT commands to provide the mean MB values for the intervention and control conditions. Here are the results:

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

New/Additional Parameters
    E_MB_C            -0.246      0.056     -4.399      0.000
    E_MB_I             0.252      0.059      4.241      0.000
```

The estimated mean MB for the control group was -0.25 ±0.11 and for the intervention group it was 0.25 ±0.11. Cohen's *d* was is the estimated covariate adjusted mean difference divided by the within-group covariate adjusted standard deviation = 0.498/0.851 = 0.59. The probability of exceptions to the rule is 0.34.

Effect of the Mediators on the Outcome

The key equation for examining the independent effects of MA, MB, and TREAT on Y using traditional regression notation is

$$\log (E[Y]) = a + b_1 MA + b_2 MB + b_3 TREAT + b_4 COV \qquad [14.1]$$

where log (E[Y]) is the log of the expected (mean) count of Y. Here is the output for the covariate estimated effects of the mediators on the outcome based on this equation:

```
MODEL RESULTS

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

Y          ON
    MA              -0.584      0.046    -12.818      0.000
    MB              -0.130      0.049     -2.654      0.008
    TREAT           -0.256      0.094     -2.713      0.007
    COV              0.178      0.047      3.810      0.000
```

The coefficient linking MA to Y was -0.58 ±0.09 (CR = 12.82, $p < 0.05$); for every one unit that MA increases, the log mean count for binge drinking over the 30 day period is predicted to decrease by -0.58 units. The coefficient linking MB to Y was -0.13 ±0.10 (CR = 2.65, $p < 0.05$); for every one unit that MB increases, the log mean count for binge drinking is predicted to decrease by -0.13 units. These coefficients are easier to interpret if translated into incident rate ratios by calculating the exponent of them. I did so in the `MODEL CONSTRAINT` commands:

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

New/Additional Parameters
    IRR_MA          0.558      0.025     21.953      0.000
    IRR_MB          0.878      0.043     20.481      0.000
```

For MA, for every one unit that it increases, the mean of Y decreases by a multiplicative factor of 0.56 ±0.05, i.e., it reduces by about (1-0.56)*100 or 44%. For MB, for every one unit that it increases, the mean of Y decreases by a multiplicative factor of 0.88 ±0.09, i.e., it reduces by about (1-0.88)*100 or 12%. As noted above, the margins of error and confidence intervals for the incident rate ratios can be asymmetrical which can be detected using bootstrapping. The bootstrapped intervals in the present case were trivially different from the above confidence intervals that I used to define the margins of error.

The IRR coefficients linking MA to Y and MB to Y do not have the property of collapsibility in negative binomial models, which is a shortcoming. For this reason, some researchers supplement the analysis using average marginal effects for the negative binomial coefficients. An advantage of the AMEs is that they take into account the presumed curvilinearity of the log function linking mean counts to the continuous predictors in the form of a single coefficient (see Chapter 12). I used the program called *Average marginal effects* on the Programs tab of my website to obtain a sense of the magnitude of the average marginal effects, although this required me moving outside of Mplus in the spirit of LISEM. Here are the results when I predict Y from MA, MB and TREAT using the R based negative binomial algorithm in the above program:

```
Average marginal effects
 factor     AME     SE       z      p   lower   upper
     ma -1.8347 0.1955 -9.3828 0.0000 -2.2180 -1.4515
     mb -0.3875 0.1452 -2.6684 0.0076 -0.6721 -0.1029
  treat -0.7309 0.2991 -2.4439 0.0145 -1.3171 -0.1447
```

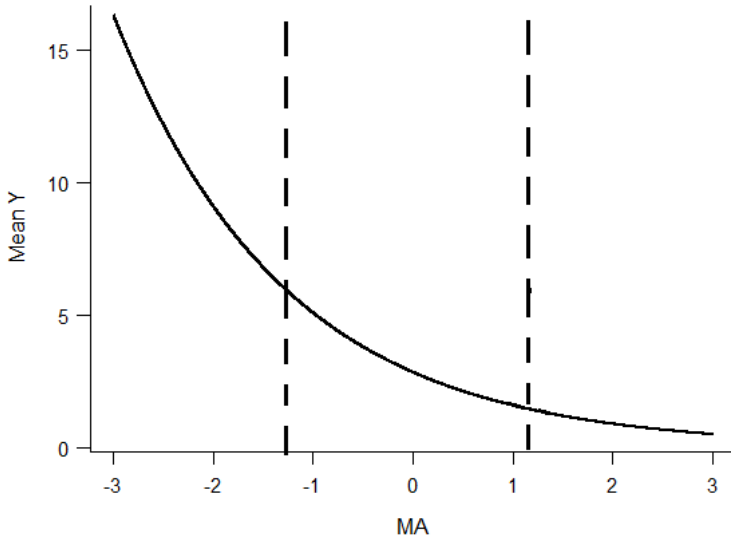
Keeping in mind the parameterizations of AMES, a one unit increase in MA is associated with a decrease in the mean number of binge drinking days of -1.83 ±0.38 controlling for the other predictors in the equation. A one unit increase in MB is associated with a decrease in the mean number of binge drinking days of -0.39 ±0.28 controlling for the other predictors in the equation. (I interpret the effect for TREAT below). These estimates have the property of collapsibility.

Meaningfulness standards can be expressed either in terms of IRRs or in terms of marginal effects depending on audience orientations. Suppose I use the marginal effects as the basis for doing so. Suppose further that program staff decide that for both MA and MB, a coefficient of -0.50 or less is meaningful. For MA, the upper limit of the confidence interval for the marginal effect is -1.45 which is lower than the standard of -0.50. The effect is therefore declared to be meaningful. If the team decides to use the same standard for MB, the upper limit of the confidence for MB is -0.10, which exceeds the meaningfulness standard. In this case, we cannot conclude the effect of MB on Y is meaningful.

For the direct effect of the treatment condition on Y independent of the two mediators, the coefficient from the primary analysis was -0.256 ±0.17 (CR = 2.71, p < 0.05). The exponent of this coefficient is 0.77 indicating that the estimated mean for the intervention condition is (1 – 0.77) * 100 = 23% lower than the mean for the control condition when MA and MB are held constant. For the average marginal effect analysis of the direct effect, the difference in the mean Y for the intervention minus control conditions was -0.73 ±0.60 controlling for MA and MB. I discuss these effects in more detail below.

Profile Analyses

It often is useful to supplement M→Y effect analyses with profile analyses. Before doing so, it often is instructive to examine the theoretical curve between each continuous mediator and the predicted Y implied by the results for Equation 14.1. Consider first the mediator MA. The observed values of MA ranged from -3 to +3. The sample means for both the covariate and MB were about 0. I used the program on my website called *Multiple curve plot* to plot the predicted Y values between MA scores of -3 and +3 for the expression exp(1.027 + -.584*X + -.130*0 + -.256*0 + .178*.08). X represents the target predictor, in this case MA. Each of the other predictors are held constant at their "typical" mean values (MB = 0, COV = 0) except for TREAT which I set equal to the control group value of 0. The expression exp calculates the predicted value such that the plotted Y is in the original Y metric rather than the log metric. The value of 1.027 is the value of the intercept for the expression taken from the Mplus output in the section called `Intercepts`. Here is the plot:



The link between MA and the levels of binge drinking is curvilinear because the negative binomial model assumes a log link function. One implication of this curve is that a one unit positive change in MA at the lower end of the MA dimension, say from the MA values of -3 to -2, should produce a much larger reduction in binge drinking than a one unit positive change at the upper end of the MA dimension, say from 2 to 3.

It turns out that the bulk of scores (80% of them) for MA occurred between -1.30 and 1.20 so the curvilinearity is less dramatic for the majority of cases. This is reflected by the curve segment between the dashed lines. The same trend occurs when I plotted MB but it was less pronounced given its smaller coefficient value appearing almost linear.

For profile analyses, I add expressions to the `MODEL CONSTRAINT` command that are of substantive interest. For example, I might compare levels of binge drinking for individuals who are at high risk of drinking versus individuals who are at low risk. I might define the former as individuals who have scores of -2 on MA, -2 on MB and who were in the control group (TREAT = 0) at the posttest versus individuals who have scores of 2 on MA and MB and who were in the intervention group. For both groups, I hold the covariate constant at its mean value, 0.08. Here are the commands I add after Line 47 of Table 14.1:

```
NEW(profile1 profile2 );
 profile1=exp(b0+p_ma*(2)+p_mb*(2)+p_treat*(1)+p_cov*(0.08)) ;
 profile2=exp(b0+p_ma*(-2)+p_mb*(-2)+p_treat*(0)+p_cov*(0.08)) ;
```

Note that for the path/regression coefficients I use the labels I assigned to them in Table 14.1. Profile 1 are low risk individuals and profile 2 are high risk individuals. Here are the results:

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

New/Additional Parameters
    PROFILE1          0.526      0.079      6.687      0.000
    PROFILE2         11.800      1.451      8.131      0.000
```

The estimated mean number of times low risk individuals engaged in binge drinking over a 30 day period was 0.53 ±0.16 whereas for high risk individuals it was 11.80 ±2.90.

Here is the case where I add `MODEL CONSTRAINT` commands that include the above two profiles but I add two additional profiles that might be of substantive interest, namely one profile (profile 3) where the MA and MB scores for individuals in the intervention group are higher up on their respective dimensions than in the high risk case (MA = -0.40 and MB = -0.40) and the second profile (profile 4) where I increase MA and MB each by half a unit to MA = .10 and MB = 0.10:

```
NEW(profile1 profile2 profile3 profile4  );
 profile1=exp(b0+p_ma*(2)+p_mb*(2)+p_treat*(1)+p_cov*(0.08)) ;
 profile2=exp(b0+p_ma*(-2)+p_mb*(-2)+p_treat*(0)+p_cov*(0.08)) ;
 profile3 = exp(b0+p_ma*(-.4)+p_mb*(-.4)+p_treat*(1)+p_cov*(0.08)) ;
 profile4 = exp(b0+p_ma*(0.1)+p_mb*(0.1)+p_treat*(1)+p_cov*(0.08)) ;
```

Here are the results for the two new profiles:

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

New/Additional Parameters

    PROFILE3           2.876      0.206     13.963      0.000
    PROFILE4           2.013      0.131     15.315      0.000
```

The estimated mean number of times individuals characterized by profile 3 engaged in binge drinking over a 30 day period was 2.88 ±0.41 whereas for individuals characterized by profile 4 it was 2.01 ±0.26. Even though profile 4 reflected an increase in both MA and MB by 0.5 (about half a standard deviation for each of them) relative to profile 3, the reduction in levels of binge drinking does not appear to be that large. This is, in part, reflective of the flattening of the curve in this region for MA (see the above Figure) and the weak effect of MB on Y.

The key point is that Mplus gives you the flexibility to pursue substantively interesting profile analyses that can complement your understanding and characterizations of the causal dynamics at play.

Omnibus Mediation

As discussed in prior chapters, my preference is to focus away from omnibus indirect effects and instead focus on the specific links within each mediational chain. A workable test of the null hypothesis for an omnibus mediation effect is the joint significance test discussed in Chapter 9. For the current example, both the T→M and M→Y links for MA as well as MB were statistically significant, indicating that the null hypothesis of zero overall mediation for both mediators can be rejected. If you want to apply the classic product coefficient method (also discussed in Chapter 9), then the relevant statistics are taken from the bootstrap analysis output. Here are the relevant results:

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

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

  Specific indirect 1
    Y
    MA
    TREAT             -0.310      0.049     -6.332      0.000

  Specific indirect 2
    Y
    MB
    TREAT             -0.065      0.026     -2.446      0.014
```

The effect of the treatment condition on the number of binge drinking days in the past month through the MA mediator was -0.31 ±0.10 (CR = 6.33, $p < 0.05$). This is the estimated log mean Y difference between the intervention group minus the control group as a function of MA. The exponent of this coefficient is 0.73 indicating that the mean for the intervention condition is (1 – 0.73) * 100 = 26% smaller than the mean for the control condition vis-a-vis MB.

The effect of the treatment condition on the number of binge drinking days in the past month through the MB mediator was -0.06 ±0.05 (CR = 2.45, $p < 0.05$). The exponent of this coefficient is 0.97 indicating that the mean for the intervention condition is (1 – 0.97) * 100 = 6% smaller than the mean for the control condition vis-a-vis MB.

Concluding Comments on Numerical Example

It is fairly straightforward to apply SEM to mediation models with count outcomes. To be sure, I did not hesitate to invoke LISEM or more traditional methods when I felt they complemented the analysis. This is an entirely reasonable approach.

One of the mediators used in this study was somewhat abstract; MA is the generalized motivation to reduce binge drinking. It is likely an intervention seeking to increase this construct would focus on a range of subtopics such as different disadvantages of engaging in binge drinking, resisting social pressures to do so, the possible negative image implications of binge drinking, as well as the positives of not binge drinking. I tend to prefer working with mediators that are less abstract because doing so often is more informative. It is likely that the non-trivial direct effect of the intervention on binge drinking independent of MA and MB represented residues of the more specific change strategies used that were intended to increase MA but that impacted binge drinking in their own right over and above MA.

MB was the felt self-efficacy to reduce binge drinking if one desires to do so. It is possible that MA and MB interact with one another to impact binge drinking. Specifically, self-efficacy should have a larger effect on binge drinking when the motivation to reduce binge drinking is high as opposed to low. When I tested for such an interaction using methods discussed in section three of this book, it was not statistically significant. It also is possible that low self-efficacy might impact motivation to reduce binge drinking (because if you do not think you can have the capacity to do something this can adversely affect your motivation to try). This dynamic would make the tested model misspecified because it omits the MB→MA link. If the MB→MA link exists it should manifest itself in the current model in the form of a non-trivial correlated disturbance between MA and MB. However, this correlation was trivial in magnitude (r = -0.08) and statistically non-significant, suggesting the omitted link likely does not exist in the current case.

I made passing reference in the example that the program staff and designers evaluated the same effect size differently when it was framed as an incident rate ratio compared to when it was framed as a more traditional mean difference. In my opinion, it is the responsibility of the program evaluator to point such disparities out to key decision makers and to explore the disparities until convergence is found. This can be challenging but framing effects should not be interfering with valid program evaluation.

ADDITIONAL CONSIDERATIONS

In this section, I consider additional analytics for SEM models with count outcomes and/or count mediators. I first discuss the use of quantile regression as a complement to more traditional count outcome modeling. I then address the use of latent variables with count indicators. Finally, I consider methods for analyzing count variables that are exogenous or that are mediators.

Quantile Regression Perspectives

When dealing with skewed distributions with potential non-linear links to other variables, some researchers use quantile regression to explore intervention and/or mediation effects on an outcome. Techniques like Poisson or negative binomial regression can fail to capture how an intervention affects the tail of a heavily skewed distribution. Quantile regression, by contrast, provides perspectives on whether the intervention reduces extreme counts which is often where the clinical or substantive value of an intervention resides. I illustrate application of this method to the binge drinking example. I assume you are familiar with the material on quantile regression from Chapter 6.

One challenge when conducting quantile regression on counts is that quantile regression assumes a continuous outcome. Count models by contrast have discrete outcomes. If the

number of counts values are substantial this can be less problematic but the presence of a large number of tied scores on the outcome also complicates the use of quantile regression. Machado and Silva (2005) proposed a solution based on the concept of **jittering** (see also Chen & Wang, 2025). Jittering is a somewhat crude but often workable solution that smooths the discrete outcome variable by adding small amounts of "jitter" to it. In essence, jittering adds a small random perturbance to each score – not enough to affect substantive results but enough to allow the quantile regression algorithms to estimate the parameters of interest. The most common way of jittering integer measures (such as a 1 to 7 rating scale) is to generate random numbers from a uniform distribution whose span is defined by the lower and upper real limit of each number. Let V be the span around the number that you wish to cover. For the score of 6, the real limits are 5.5 to 6.5, so the span is 1. For the score of 1, the real limits are 0.5 to 1.5, so again the span is 1. Thus, V = 1 for each score on the variable. Random noise is then added to a given score using a uniform distribution between –V/2 and +V/2. This is called **symmetric jittering**. The span is called the **dequantization value**. It assumes that people with a score of 5 actually scored somewhere between 4.5 and 5.5 on the underlying continuous dimension and that these individuals are uniformly distributed across this span.

For count scenarios one typically uses instead an algorithm called **to-the-right jittering**. In this case, random noise is added to a score based on a uniform distribution from 0 to V where V is commonly set to 1 instead of –V/2 to +V/2. This approach is mathematically preferred over symmetric jittering for count data because it preserves a monotonic one-to-one relationship between the continuous quantiles of the jittered variable and the discrete quantiles of the original count variable (Machado & Santos-Silva, 2005).

I must admit that it feels a bit perverse to purposely add random noise to data. But the underlying assumptions of jittering often are reasonable and if its use ultimately allows us to apply more powerful methods of analysis, then this is desirable. The work of Machado and Santos-Silva (2005) makes a reasonable case for the use of the approach with counts.

When applying quantile regression to a count-based RET, I usually conduct an initial analysis focusing on the outcome and a single binary predictor, namely the treatment condition dummy variable. The result then informs me if the intervention is having larger or smaller effects on the outcome at different points in the distribution. I decided to focus on the middle versus upper end of the distribution in the binge drinking example, so I conducted the analysis using the *Quantile regression* program on my website for the 50th quantile (the median), the 75th quantile, and the 90th quantile.

I can use either a conditional quantile regression model or an unconditional model as discussed in Chapter 6. Because there is only a single predictor and no other covariates, the two methods should yield similar results. However, the two approaches estimate slightly different parameters and in this case I give precedence to the conditional quantile regression model as implemented in the R package *quantreg* and operationalized on my website. The

reason I prefer it is because with random assignment to the two treatment conditions, both the intervention and control groups represent the same population at baseline and they should be balanced on other covariates. The conditional model calculates the exact mathematical sample difference between the group quantiles whereas the unconditional quantile regression uses a linear approximation based on the density of the outcome. This approximation can add some estimation noise. The conditional model with a single predictor tends to be robust and also avoids the need for kernel density estimations required by the unconditional methods. Here is the (edited) output from my webpage:

```
Model coefficients for q=0.5
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.9388     0.2621 11.2146   0.0000
treat       -0.9901     0.3131 -3.1627   0.0017

Model coefficients for q=0.75
              Value Std. Error t value Pr(>|t|)
(Intercept)  6.3797     0.4459 14.3086        0
treat       -2.8892     0.4992 -5.7878        0

Model coefficients for q=0.9
              Value Std. Error t value Pr(>|t|)
(Intercept) 10.1560     1.0232  9.9256   0.0000
treat       -3.8309     1.2419 -3.0848   0.0022
```

For each quantile, the intercept represents the estimated quantile value for the control group. The estimated median instances of binge drinking for the control group was 2.94 ±0.52. The median for the intervention group was -0.99 ±0.62 units lower than this value, a difference that was statistically significant (CR = 3.16, $p < 0.05$). The estimated instances of binge drinking at the 75th quantile for the control group was 6.38 ±0.89. The corresponding 75th quantile for the intervention group was -2.89 ±1.00 units lower than this value, a difference that was statistically significant (CR = 5.79, $p < 0.05$). Finally, the estimated instances of binge drinking at the 90th quantile for the control group was 10.16 ±2.05. The 90th quantile for the intervention group was -3.83 ±2.48 units lower than this value, a difference that was statistically significant (CR = 3.09, $p < 0.05$).

The general trend in these results is that the intervention tends to become increasingly effective at the higher quantiles of the outcome. The quantile regression program also provides tests of the null hypothesis that the coefficients for each pair of quantile coefficients is non-zero. Here are the results of these tests:

```
Sig test for q=0.5 vs. q=0.75 coeff difference
      df1 df2 F statistic p value
treat   1 999     21.4822       0
```

```
Sig test for q=0.5 vs. q=0.9 coeff difference
      df1 df2 F statistic p value
treat   1 999      5.8432  0.0158

Sig test for q=0.75 vs. q=0.9 coeff difference
      df1 df2 F statistic p value
treat   1 999      0.8194  0.3656
```

The tests indicate that the contrasts are significantly different from each other with the exception of the 75th quantile versus the 90th quantile.

I repeated these analyses using the unconditional quantile model and the results were comparable to those of the conditional model. With jittering, it is good practice to replicate analyses using different random seeds, somewhere between 5 and 20 imputations. When I did so, none of the major conclusions from the above changed in any of the analyses.

Quantile regression adds insights to the more traditional negative binomial analysis. The negative binomial model reports the incidence rate ratio which is the estimated mean Y of the intervention group divided by the mean Y of the control group. The quantile regression extends the analysis not only to intervention effects in the middle of the distribution but, in the current case, to the evaluation of quantiles in the upper end of the distribution using additive as opposed to multiplicative perspectives. I have more to say about this shortly.

I repeated the quantile regression analysis predicting MA and MB, separately, from TREAT but did not find the same trend as above. Rather, the effect of the treatment condition on each mediator was relatively uniform. However, when I performed an unconditional quantile analysis regressing Y onto MA, MB, TREAT and COV simultaneously, an interesting dynamic emerged. Here is the output from the analysis using the *Quantile regression* program:

```
UNCONDITIONAL QUANTILE REGRESSION

Unconditional results for q=0.5
               Coef   Lower   Upper Std.Error  Z value P Value
(Intercept)  2.6011  2.2466  2.9797    0.1870  13.9092  0.0000
ma          -1.2348 -1.4532 -1.0069    0.1139 -10.8455  0.0000
mb          -0.2891 -0.5381 -0.0244    0.1311  -2.2064  0.0274
treat       -0.3734 -0.8943  0.1551    0.2677  -1.3947  0.1631
cov          0.3395  0.0914  0.6039    0.1307   2.5968  0.0094

Unconditional results for q=0.75
               Coef   Lower   Upper Std.Error Z value P Value
(Intercept)  5.4788  4.6145  6.3378    0.4396 12.4628  0.0000
ma          -2.8429 -3.3835 -2.2507    0.2890 -9.8382  0.0000
mb          -0.6661 -1.2163 -0.0982    0.2852 -2.3352  0.0195
treat       -1.5139 -2.7225 -0.3665    0.6010 -2.5187  0.0118
cov          0.6575  0.0237  1.2752    0.3193  2.0593  0.0395
```

```
Unconditional results for q=0.90
               Coef   Lower   Upper Std.Error Z value P Value
(Intercept)  9.0579  7.2408 10.8542    0.9218  9.8262  0.0000
ma          -5.6837 -7.2767 -4.0634    0.8197 -6.9335  0.0000
mb          -1.1425 -2.3438  0.1457    0.6351 -1.7990  0.0720
treat       -1.2873 -3.6978  1.1151    1.2278 -1.0484  0.2944
cov          1.4632  0.2420  2.7548    0.6410  2.2826  0.0225
```

Note that for both MA and MB, the coefficient reflecting the impact of the mediator on the outcome is stronger at the higher quantiles. For example, the MA coefficient at the median is -1.23, at the 75th quantile it is -2.84, and at the 90th quantile it is -5.68. Stated more generally, MA is disproportionately effective for the heaviest drinkers. It does not just uniformly shift the distribution; it also compresses the upper tail of the distribution. This helps explain the earlier pattern for the effect of the intervention on binge drinking.

Part of the reason the median coefficient is smaller than the 75th and 90th quantile coefficients is structural. Students at or below the median have relative few binge drinking days to begin with, so they hit a "floor" of zero quickly. In the 90th quantile, we do not encounter the floor so MA can better reveal its behavioral strength.

Recall that the incidence rate ratio for MA in the negative binomial model was 0.56; a one unit increase in motivation multiplies the expected number of binge drinking days by 0.56 (a 44% reduction). Mathematically, a 44% reduction differs depending on where a student starts on the binge drinking dimension, either at the low end of the distribution, in the middle of the distribution, or at the high end of the distribution. For a student who initially engages in binge drinking, say, 4 times a month, a 44% reduction translates into 1.76 fewer binge drinking days. For a student who initially engages in binge drinking 20 times a month, a 44% reduction translates into a reduction of 11.2 fewer drinking days. The negative binomial model is sensitive to this dynamic but note that it forces the percent reduction, in this case 44%, to be the same across the entire Y dimension. The quantile regression relaxes this restriction, allowing for more flexible effect heterogeneity.

Interestingly, if instead of analyzing Y in the quantile regression one analyzes log(Y), this changes the interpretation of the quantile coefficients from absolute differences on the original scale (additive shifts) to multiplicative effects or proportional changes more akin to the negative binomial model (Koenker, 2005). When I did so for the current data, the coefficients were comparable across the different quantiles and similar to the negative binomial result of 0.56 for MA.[2] In some ways, the quantile regression can be viewed as a methodological check on the negative binomial parametric assumptions while also providing a different (additive rather than multiplicative) vantage point on the data.

[2] This also was true of MB. Without jittering, log based analysis with count data falls apart if there are zero values in the outcome because the log of zero is undefined. Also, if the units of analysis have different exposure opportunities, one will want to include exposure offsets when logging.

I do not want to get side tracked on the intricacies of the quantile regression approach for the current data. I simply want to draw attention to the potential use of quantile regression for analyzing count data coupled with the use of the Machado and Santos-Silva jittering method. Quantile regression has the advantage of being semi-parametric rather than parametric in nature, it tends to be more outlier resistant on the outcome relative to the negative binomial model, and it offers different vantage points on the causal dynamics while also allowing for effect heterogeneity.

Latent Variables with Count Indicators

Sometimes RETs have "count" outcomes that constitute multiple indicators of a latent variable, i.e. the outcome is a latent variable with count indicators. In such cases, one programs Mplus as above but creates a latent variable with a reference indicator per standard latent variable practice. In this scenario, the latent variable is continuous, not a count. This is because the unobserved latent variable representing the shared variance among the indicators is conceptualized as a normally distributed continuous construct. Given this, you model its structural relationships in the same way as any other latent continuous outcome.

Importantly, with count indicators you still treat the indicators as counts by specifying them as such using the `COUNT` command. This tells Mplus to use an appropriate link function (e.g., negative binomial) for the measurement model. Unlike standard continuous latent indicators, count indicators do not have an estimated error/residual variance parameter because the variance of a count distribution is a function of its mean. However, the latent variable itself will have a disturbance parameter when it is an outcome in a structural path.

The metric of the latent variable is defined by fixing the factor loading of the one of the count indicators to 1.0. The intercept of that indicator is fixed to 0. This forces the latent variable to take on the metric of that indicator's log-scale. Because the scale of the observed count indicator maps directly onto the factor via the chosen count distribution (typically a negative binomial or Poisson distribution), the latent construct is on a log scale. This means the path coefficients for the predictors of the latent variable are interpreted as you would in a standard Poisson or negative binomial regression. The coefficient per se reflects how the log of the latent variable changes given a one unit increase in the predictor holding the other predictors constant. The exponent of the coefficients is the incident rate ratio or multiplicative factor for the predictor holding constant the other predictors.

If the indicators of a latent variable are a mix of count and continuous indicators, then additional challenges arise. In this case, the measurement model essentially states that a unit increase in the latent variable produces a linear change in the continuous indicators but an exponential change in the count marker variable. Sometimes this is justifiable but other times it is questionable which then creates a misspecified measurement model. Some methodologists

log-transform the continuous indicators before executing the model so that all indicators behave linearly with respect to the log-scaled latent variable.

As an example, suppose the latent variable is the addiction to on-line shopping and it has three indicators (a) the number of packages delivered (a count indicator used as the marker variable with a negative binomial distribution), (b) monthly credit card debt in dollars (a continuous indicator), and (c) hours spent browsing shopping apps (a continuous indicator). The latent variable has a log link given the count indicator is its marker variable. Suppose a person moves from a latent addiction score of 1 to an addiction score of 2. This means the number of packages delivered will increase exponentially but the credit card debt and browsing time will increase linearly. This may be unrealistic. Rather, one might expect all three indicators to increase exponentially. By calculating the log of the continuous indicators, one brings comparability into line. All indicators now reflect exponential growth driven by the latent trait.

Count Exogenous Predictors and Mediators

In Mplus, you can only declare a variable as a count when the variable is endogenous. If your count variable is exogenous, Mplus will treat it as it would any other quantitative variable. This is not unreasonable with the only potential problem being if the exogenous count variable is formally brought into the model by declaring in Mplus syntax to estimate its mean, its variance or its correlation with some other variable. Otherwise, Mplus treats it as fixed which renders it distribution irrelevant estimation wise.

If the count variable is a mediator, M, but is declared as a count in the `COUNT` command, Mplus operates on it differently in the context of the mediational chain. When M is predicted by, say, the intervention dummy variable, Mplus treats it as a count variable and applies the scaling you have specified for it (Poisson or negative binomial). When M predicts the outcome, Y, Mplus treats M as a continuous/linear predictor. It assumes that the observed count values (0, 1, 2, 3...) are meaningful linear steps.

Mplus disables the `MODEL INDIRECT` command when a count variable is a mediator. This is because one of the paths is on a log scale and the other is on a linear scale, which complicates the meaning of the product of the two coefficients. Instead you must rely on the joint significance test to evaluate null hypothesis tests of mediation in such cases. If the mean of your count variable is large and not heavily zero-inflated, it often behaves much like a continuous variable. In such cases, some researchers choose not to declare it in the `COUNT` command, in which case you can use the `MODEL INDIRECT` command.

CONCLUDING COMMENTS

In some ways, the term "count regression" is a misnomer because the methods described in this chapter can be applied to any discrete, multi-category outcome. A better term probably

is discrete regression modeling. Counts are one of the most common discrete outcome variables in the social and health sciences, so it is only natural that discrete regression models have come to be called count regression models. But technically, they apply to far more than counts.

Count data can approximate a wide range of distributions. They can approximate normal distributions, Poisson distributions, negative binomial distributions, zero inflated distributions, and hurdle-like distributions, among others. When choosing a model, one must think about how the predictors in the model are related to the mean outcome as one moves across the scores on those predictors. If the relationship is linear, then traditional count models are potentially problematic because these models assume the predictors are linearly related to the log of the mean count. For linear relationships, some form of robust regression can be used (e.g., Huber-White based robust regression, bootstrapping) or some form of quantile regression might be appropriate. For cases where some of the predictors are linearly related to the mean count, while others are linearly related to the log mean count, then robust regression with non-linear modifications (e.g., predictor polynomials) might be feasible of perhaps some form of the generalized additive model discussed in Chapter 15. The general point is that one needs to think long and hard about the distribution of Y as well as its functional relationship with its presumed determinants when choosing an appropriate model. In my experience, negative binomial and hurdle models generally accommodate data better than the other count regression models, but, of course, much depends on the substantive area one is working in. I frequently hear researchers say "if you have count data, you should conduct a Poisson regression," without due appreciation of the more subtle modeling dynamics that must be taken into account.

For rates, discrete regression models with offsets are an interesting way to approach such analyses, but, technically, the approach is mainly appropriate only for Poisson modeling, which is rather limiting. Viable analytic alternatives include fractional logit/probit modeling and generalized additive models.