5

Statistical Fundamentals: Regression

Regression analysis is the hydrogen bomb of the statistics arsenal

- CHARLES WHEELAMORTURE

INTRODUCTION

In this chapter, I review regression frameworks that form the backbone of RET analysis. In chapter 6, I discuss a wide range of additional statistical issues that complement the regression fundamentals discussed here. In chapter 7, I introduce structural equation modeling (SEM) and in Chapter 8 I consider non-traditional SEM. As a collective, the chapters provide the statistical fundamentals for the analysis of RETs. I assume you are somewhat familiar with the topics in the current chapter; my exposition is more conceptual than mathematical. Most of the sections can be read on a stand alone basis and you can jump around the chapter as dictated by your interests.

THE BASICS OF LINEAR REGRESSION

The analysis of many RETs is grounded in regression modeling. Probably the most popular regression method used to analyze RETs is fixed predictor ordinary least squares (OLS) regression. It is used for both mediation and moderation analysis, although as I will show in future chapters, more modern methods of analysis are now available. Nevertheless, understanding the basics of linear regression is essential for RET analysis.

Linear regression uses outcomes that are continuous or quantitative-discrete with many values. In bivariate regression that examines the relationship between X and Y, the linear model has the form

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

where α is the intercept, β is the unstandardized regression coefficient, and ε is an error term, also called a disturbance term.[1] With multiple predictors, the equation has the form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta X_k + \varepsilon$$

where *k* is the number of predictors and all other terms are as previously defined. The

[1] The term *error* often is reserved when referencing a population parameter and residual for the disparity between predicted and observed score in samples. However, term use varies by field making it confusing to be consistent. I sometimes shift terminology depending on the tradition of the area I am discussing.

equations sometimes include a subscript *i* for the X, Y and ε terms to reflect the fact that there is a separate score for each individual. I omit the subscript and assume it is implicit.

One way of thinking about linear regression is in terms of conditional means. Suppose the outcome variable is the amount of time an adolescent spends with his or her mother in a week and the predictor is the age of the adolescent, with age ranging from 12 to 17 years. I hypothesize that older adolescents spend less time with their mothers. I might segregate the sample into the different age groups and calculate the mean amount of time spent with the mother, as follows:

Age	Mean Hours Spent Together
12	32
13	30
14	28
15	26
16	24
17	22

The mean number of hours spent together conditioned on age being 12 is 32 hours. The mean number of hours spent together conditioned on age being 13 is 30 hours. And so on. Note that for these data, the means change as a linear function of age: For every one unit that age increases, the mean amount of time spent together decreases by 2 hours. A regression analysis of these data would yield an unstandardized regression coefficient of -2.0. This coefficient provides perspectives on how the mean of Y changes across X values. The changes are presumed to be linear. Regression analysis, at essence, analyzes conditional means and how the means vary as a function of predictor values, although it provides additional perspectives on data in terms of individual variability about the conditional means.

In order to develop uses of regression analysis in RETs, I use an example that compares two types of programs to increase adherence to medical regimens for people living with HIV. Each program addresses the same three facets, (1) teaching people strategies to cope with the side effects of the medications, (2) teaching people how to strengthen social support for protocol adherence, and (3) increasing perceptions of the importance of adhering to the protocol. One program uses passive learning techniques to address each mediator such that participants read engaging and informative articles on each topic. The second program uses the same methods but augmented with skills-based and active learning strategies, such as role playing, writing essays, and participants explaining materials to other participants. With an active control group, the RET has three conditions. I refer to them as (a) the education group, (b) the education plus skills group, and (c) the control group.

Figure 5.1 presents the conceptual logic model for the program using an influence diagram. The figure includes a path directly from the treatment condition to the adherence outcome to reflect the fact that the programs may impact adherence through mediators not specifically targeted, such as program satisfaction (see Chapter 2). The adherence outcome was measured as the percent of protocol compliance over a one month period. It ranged from 0 to 100 and was measured 3 months after program completion. The mediators were measured 1 month post-program completion. A clinician rating of the quality of the patient's social support was based on a structured interview by the clinician. Scores ranged from 0 to 10 (0 = no support, 3 = slight support, 6 = moderate support, 9 = strong support). The clinician could make finer gradations across the 0 to 10 metric by assigning decimals. Clinicians received training about how to use the scale, which had concrete behavioral anchors for each major scale category. A comparable 0 to 10 clinician-based rating was used for the quality of side effect coping strategies used, and a 0 to 10 self-report of adherence importance was obtained from each individual (0 = not at all important, 3 = slightly important, 6 = moderately important, 9 = very important). To make matters simpler for purposes of exposition, I use a three-group posttest only control group design so there are no baselines assessments of any variables.

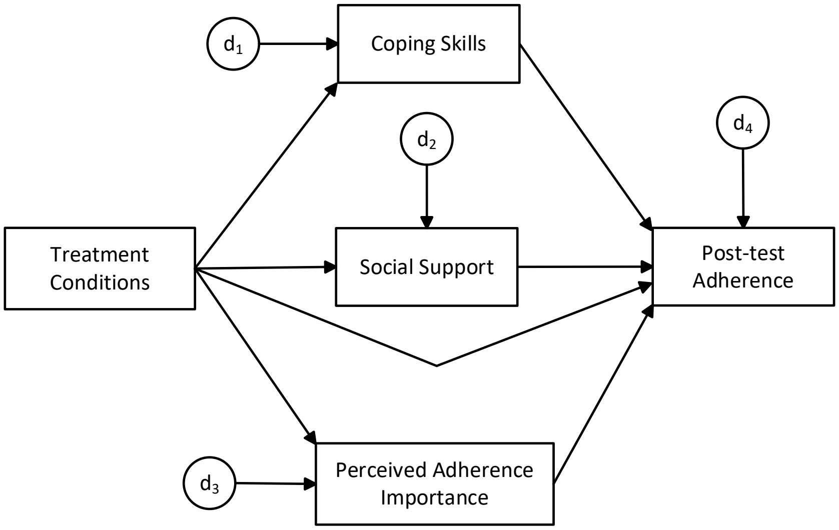

FIGURE 5.1. Logic model for adherence intervention

Linear Regression with Nominal Predictors

One type of regression used in RETs is when a mediator, M, is regressed onto one or more variables representing the treatment conditions to which people are assigned. I illustrate this for the model in Figure 1 by regressing social support, onto the treatment conditions. Because the treatment condition is nominal, it is represented using dummy variables. A **dummy variable** is a variable with a set of scores assigned by the analyst to capture group membership. With **dummy coding**, only 1s and 0s are assigned. For the dummy variable D_E, I assign 1s to all individuals who participated in the education program and zeros to everyone else. For the dummy variable D_{ES}, I assign 1s to all individuals who participated in the education plus skills program and zeros to everyone else. For the dummy variable, D_C, I assign 1s to all individuals assigned to the control condition and zeros to everyone else. When modeling the data, I can only use two of the three dummy variables in the regression equation because the third one is completely redundant with the other two; if I know whether a person participated in the education program and also whether the person participated in the education plus skills program, then, by definition, I know if the person was in the control group. The group whose dummy variable is omitted from the regression equation is called the **reference group**. More generally, for a nominal variable with *k* levels, we use *k*-1 dummy variables in the equation. The choice of which group serves as the reference group is decided on substantive grounds or it is arbitrary, as I elaborate below. The social support measure reflects a mediator of program effects on the outcome. I want to determine the effects of the programs on it, so it is the dependent variable. The dummy variables for the program conditions are thought of as a collective representing this nominal variable as a whole.

For dummy variables that use dummy coding, the regression coefficient associated with a given dummy variable is the predicted mean difference on the outcome (social support) between the group scored 1 on the dummy variable and the reference group, holding all other predictors in the equation constant (in this case, there are no other predictors). For example, if I use D_E and D_{ES} as predictors, the regression coefficient for D_E equals the predicted posttest mean social support for people who participated in the education program minus the posttest mean social support for people in the control group.

In the current example, it makes substantive sense to use the control group as the reference group. However, often researchers will re-analyze the data by changing the reference group to explore other contrasts. For example, if I re-analyze the data using D_{ES} and D_C as predictors, the regression coefficient for D_{ES} will equal the posttest mean social support for those in the education and skills program minus the corresponding mean for the education program.

The population level equation in this analysis takes the following form:

$SS_{POST} = \alpha + \beta_1 D_E + \beta_2 D_{ES} + \varepsilon$

SS_{POST} is social support measured at the posttest. Suppose I analyze the sample data and obtain the following results:

$$SS_{POST} = 3.0 + 2.0\ D_E + 3.0\ D_{ES} \qquad [5.1]$$

The intercept is the posttest social support mean when all the predictors equal 0. When both D_E and D_{ES} equal zero, this defines the control group, so the intercept is the posttest mean level of support for the control group. It is 3.0 on its 0 to 10 metric and reflects a somewhat low level of social support. The coefficient for D_E is 2.0, which is the posttest mean support difference between those in the education program and those in the control group. The education program increases social support, on average, by 2.0 units on the 0 to 10 social support metric. The coefficient for D_{ES} is 3.0, which is the posttest mean support difference between those in the education plus skills program and those in the control group. Of course, we are interested in the t tests, p values, and confidence intervals for these parameter estimates, but I do not report them here.

If I re-analyze the data by changing the reference group to the education program, the resulting regression equation is

$SS_{POST} = 5.0 + -2.0\ D_C + 1.0\ D_{ES}$

In this case, a score of 0 on D_C coupled with a score of 0 on D_{ES} reflects the participants in the education program. Their posttest mean support is the intercept, which equals 5.0. The coefficient for D_C is -2.0, which is the mirror image of the coefficient for D_E in the first analysis. This information is redundant because in the first analysis, we subtracted D_C from D_E and in this analysis we subtracted D_E from D_C. The coefficient for D_{ES} is 1.0, indicating that the posttest mean support for those in the education plus skills program is 1.0 unit higher than the education group. This reflects a different contrast than the contrasts in the first analysis. By conducting these two regressions, I obtain information on the three contrasts that are of substantive interest, namely D_E versus D_C, D_{ES} versus D_C and D_{ES} versus D_E. The relevant means and confidence intervals for each group separately are obtained from the intercepts. I only derived two of these group means because I calculated two intercepts. The third mean of interest can be calculated by re-analyzing the data yet a third time but making the education plus skills program the reference group. The intercept for this analysis will provide us with the predicted posttest mean for this group and its associated confidence interval. Table 5.1 summarizes the key results extracted from the three regression analyses.

Table 5.1: Contrasts Using Dummy Variables

Contrast	Mean 1	Mean 2	Mean Difference	t ratio
ES vs. C	6.0	3.0	3.0	6.00
E vs. C	5.0	3.0	2.0	4.00
ES vs. E	6.0	5.0	1.0	2.00

Linear Regression with Quantitative Predictors

Another type of regression analysis used in RETs is when the outcome is regressed onto the mediators. To illustrate, I again use the adherence example. I regress the post program adherence measure onto all the variables with arrows pointing directly to adherence in Figure 5.1. The predictors are the three mediators and the treatment condition dummy variables. Here is the relevant population equation:

$$AD_{POST} = \alpha + \beta_1 \text{ CSE} + \beta_2 \text{ SS} + \beta_3 \text{ Imp} + \beta_4 \text{ D}_E + \beta_5 \text{ D}_{ES} + \varepsilon \qquad [5.2]$$

and the results from the analysis of sample data are:

$$AD_{POST} = 25.0 + 1.0 \text{ CSE} + 1.5 \text{ SS} + 2.0 \text{ Imp} + 0.1 \text{ D}_E + 0.2 \text{ D}_{ES} \qquad [5.3]$$

A regression coefficient for a quantitative predictor reflects the number of units the posttest mean of the outcome is predicted to change given a one unit increase in the value of the predictor, holding all other predictors in the equation constant. The coefficient for coping skills for side effects was 1.0. This means that for every one unit that the coping skills increases, the mean adherence is predicted to increase by 1.0 (percentage) units, holding constant all other variables in the equation. The coefficient for the social support mediator was 1.5. This means that for every one unit that social support increases, the mean adherence is predicted to increase by 1.50 (percentage) units, holding constant all other variables in the equation. The coefficient for the perceived importance of adhering mediator was 2.0. This means that for every one unit that the perceived importance of adhering to the protocol increases, the mean adherence is predicted to increase by 1.50 (percentage) units, holding constant all other variables in the equation. The magnitude and statistical significance of these coefficients are important because they tell us if the assumptions that the program designers made about the relevance of these mediators are viable, per my discussion in Chapter 1.

Combining Results of Two Regression Analyses

Sometimes when analyzing RETs we combine results from two separate regression analyses to gain perspectives of substantive interest. From Equation 5.1, we found that the effect of the education program relative to the control group was to raise social support, on average, by two units, which is the value of the coefficient for D_E. In a separate regression analysis using Equation 5.3, we found that for every one unit that social support increases, the mean medication adherence increased by 1.5 units, holding constant the other predictors in the equation. It follows from this that the two unit increase in social support vis-a-vis the education program should increase medication adherence, on average, by (2.0)(1.5) = 3.0 (percentage) units, holding constant all other predictors in Equation 5.3. This latter value represents the estimated effect of the education program on the adherence outcome *through the social support mediational chain*. Generically, the product of the two coefficients is called the **indirect effect** of the program on the outcome through the social support mediator. This indirect effect is derived by combining information from the two regression equations, namely (1) the estimated effect of the education program on social support multiplied by (2) the estimated effect of social support on adherence. When analyzing RETs, we often derive new parameters of substantive interest, such as the magnitude of an indirect effect through a given mediational chain, by combining parameters from different equations.

Purposely Omitting Variables from a Regression Analysis

Another practice sometimes used when analyzing RETs is to strategically omit predictors from a regression equation in order to isolate a parameter of substantive interest. For example, I might want to estimate the overall effect of the education program on the adherence outcome and the overall effect of the education plus skills program on the adherence outcome. These effects are called the **total effects** of each program and can be estimated using the following equation:

$$AD_{POST} = \alpha + \beta_1 D_E + \beta_2 D_{ES} + \varepsilon \qquad [5.4]$$

β_1 is the mean adherence difference between the education group and the control group, and β_2 is the mean adherence difference between the education plus skills group and the control group. Note that even though the mediators are determinants of adherence, I have purposely omitted them from the equation in order to isolate the overall effect of the programs on the outcome.

Taken together with the approach in the previous section, these examples illustrate that we can derive parameter estimates of substantive interest in RETs by strategically including or excluding predictors in a linear equation. We include variables we want to hold constant or control when estimating the parameter of interest and we exclude variables that we do not want to hold constant or control.[2]

Consider yet another example that makes use of Equation 5.3 that analyzes the effects of mediators on the outcome. I repeat the equation here for convenience:

$$AD_{POST} = 25.0 + 1.0\ CSE + 1.5\ SS + 2.0\ Imp + 0.1\ D_E + 0.2\ D_{ES}$$

Consider the coefficient for D_E which has a value of 0.1. This is the estimated effect of the education program minus the control group on post-test adherence *independent of the three targeted mediators*. This is because the three mediators are statistically held constant vis-à-vis their inclusion in the prediction equation. If the coefficient for D_E is near-zero or weak (which it is in the present case), this suggests that the primary impact of the education only program, if any, is through one or more of the program-targeted mediators. If the D_E coefficient is large and non-trivial, then this suggests the education program has effects on the outcome through mechanisms not directly targeted by the program, such as participant satisfaction with the program or via relationships the participant has with program staff and leadership, per my discussion in Chapter 2. The coefficient for D_{ES} is the corresponding parameter for the education plus skills program. In the present case, the estimated coefficients for D_E and D_{ES} were both small and nonsignificant, suggesting any effects of the programs on adherence were through the mediators that the programs targeted.

To summarize, I can strategically define equations that allow me to derive indices of (a) the indirect effects of a program on the outcome through a given mediator, (b) the direct effect of a program on the outcome independent of the target mediators, (c) the effect of a mediator on the outcome independent of the other mediators, and (d) the total effect of the program on the outcome. I nuance these ideas more in future chapters. The take-away for now is how to interpret coefficients in a linear equation when analyzing RETs and an appreciation for combining coefficients from different equations as well as adding/removing covariates to answer questions of substantive interest.

Population Assumptions for Linear Regression

Using sample data, researchers estimate population parameters of a regression model and calculate confidence intervals and significance tests for those estimates. For results to be valid, traditional OLS regression makes assumptions about the population. I

[2] In later chapters, I introduce full information SEM that estimates indirect and total effects not in separate regressions but in a multivariate, single estimation approach. The current strategy represents what is called limited information SEM.

state key assumptions here but do so somewhat informally in the interest of conveying an intuitive sense of them.[3] The assumptions are:

1. ***Error scores are normally distributed***: This assumption holds that for any given predictor profile defined the regression equation, the population ε scores for individuals characterized by that profile are normally distributed. For example, for Equation 5.2, if I segregated every individual in the population with the profile of CSE = 9, SS = 9, IMP = 9, D_E = 0 and D_{ES} = 0, and I examined the error scores for these individuals, the error scores would be normally distributed. If I segregated every individual in the population with the profile of CSE = 1, SS = 1, IMP = 1, D_E = 0 and D_{ES} = 0, the error scores for this profile also would be normally distributed.

2. ***Error score means are zero***. This assumption holds that for any given predictor profile, the mean of the error scores for individuals characterized by that profile is zero. The logic behind this assumption is that the error scores reflect factors not included in the equation that impact the outcome. Some of these factors will push adherence scores upwards (positive errors) and others will push adherence scores downward (negative errors). The net effect will be a mean of zero across individuals in that profile, as the positive errors cancel out the negative errors when I calculate the mean.

3. ***Error score variances are homogeneous***. This assumption holds that for any given predictor profile, the variance of the error scores for individuals characterized by that profile will be the same as the variance of the error scores for individuals characterized by any other given profile.

4. ***Predictor-error correlations are zero***: This assumption holds that across all individuals in the population, each predictor in the equation is uncorrelated with the population error scores. Stated another way, if I correlate scores on predictor X with people's error scores, the correlation will be zero. As noted above, the error scores are assumed to represent random influences of variables external to the equation on the outcome. Sometimes these external influences push outcome scores upward and sometimes downward. These random influences are assumed to impact the outcome in the same way no matter what a person's score is on a predictor, i.e., they are assumed to operate like random noise in the system specified by the equation. If I correlate a set of random numbers with a predictor X, the correlation should be zero. In more technical terms, the assumption is that the predictor values and the error scores are independent.

[3] When stating the assumptions, I use the term "error scores" in place of "disturbance scores" because this is the terminology used in most regression texts.

5. ***Errors are independent***: This assumption holds that the population error scores are independent. Knowing an error score for one person does not allow us to predict another person's error score.

Note that all of these assumptions focus on the population ε scores; no assumptions are made about the distributions of the predictors per se. This is why dummy variables can be analyzed effectively in fixed predictor regression contexts.

It is not uncommon for some of the above assumptions to be violated. When this is the case, significance tests and confidence intervals can be biased. Some researchers believe that OLS regression is highly robust to assumption violations, but there is controversy about such assertions. For good discussions of robustness, see Wilcox (2017) and Wilcox & Rousselet (2018). As examples of robustness complications, when comparing group means with dummy variables in the presence of non-normal but identical error score distributions in the groups, Type I error rates generally are well maintained with OLS regression. However, if the error score distributions differ non-trivially between the groups, Type I error rates and/or confidence interval coverage can be compromised. As another example, when error scores are normally distributed but the variances are heterogeneous, the statistical power of OLS regression can suffer, sometimes considerably. Power also can be adversely affected when error score variances are homogenous but their error distributions are heavy tailed. The general point is that one should not take OLS assumptions lightly, a topic I elaborate in the next section.

Robust Regression: When Assumptions are Violated

A popular strategy for dealing with assumption violations of normality and variance heterogeneity in linear models is known as **bootstrapping**. Bootstrapping is a non-parametric empirical approach to estimating sampling distributions, standard errors, confidence intervals, and p values. Consider the case of a correlation coefficient. The standard error and confidence interval for a correlation coefficient are not mathematically derivable. Suppose one obtains a random sample of 500 cases from a population and calculates a correlation coefficient in the sample. The idea behind bootstrapping is to use the sample data to mimic population data and then, with access to the "population" data, estimate the standard error of the sampling distribution and the corresponding confidence interval and p value for the parameter, in this case a correlation. One selects a random sample of 150 cases from the sample data of 150 cases using sampling *with replacement*. The new sample is called a bootstrap sample or a **bootstrap replicate**. By using sampling with replacement, the initial sample of scores mimics an infinitely large population whose scores are distributed the same as the sample data. The bootstrap replicate represents a random sample from that population. Using high speed computers, one generates thousands of bootstrap samples/replicates and treats the resulting statistic calculated in each sample as a sampling distribution for that parameter. One then empirically derives an estimate of the standard error by calculating the standard deviation of the sampling distribution. These empirically derived standard errors can be used to calculate confidence intervals or one can use the estimated sampling distribution per se to calculate quantile values that map onto the desired confidence interval, such as the 0.025 quantile and 0.975 quantile of the sampling distribution. A powerful feature of bootstrapping is that the logic can be used for any statistic (a median, a regression coefficient, a correlation coefficient) to estimate standard errors, confidence intervals, and p values when assumption violations question traditional OLS statistical inference.

There are different strategies one can use to estimate p values and confidence intervals from bootstrap replicates. These include the percentile bootstrap method, the bootstrap t method, and the bias-corrected method, among others. Wilcox (2012, 2017) discusses differences between the approaches. Bootstrapping does not always yield workable results (Hesterberg, 2015). For example, bootstrap confidence intervals tend to be too narrow for small samples and must be used with caution in such cases. Wilcox (2012, 2017) also describes scenarios where bootstrap methods tend to work well and where they fail. I make use of bootstrapping in later chapters.

Another popular approach to assumption violations is the use of a **Huber-White robust estimator** (see Wilcox, 2017, for a description of this method). This approach estimates coefficients using ordinary least squares regression but adjusts standard errors to accommodate variance heterogeneity and many types of non-normality. There are different forms of this estimator, with type HC3 typically being best for smaller sample sizes (Long & Ervin, 2000). Mplus uses HC0, which is best for N > 250.

A common strategy used to address possible assumption violations is to first apply a statistical test for the presence of violations, such as a test of variance heterogeneity or a test of non-normal error scores. If the preliminary test yields a statistically significant result, one shifts to using a robust method; if not, one uses traditional OLS methods of analysis. This two-step approach is problematic for several reasons. First, many preliminary tests lack statistical power. Without large sample sizes, they can yield non-significant results even when the violation is problematic (Wilcox, Charlin & Thompson, 1986; Wilcox, 2003). Second, the critical issue for robustness is not whether an assumption is violated, but rather the nature and *degree* of violation. We know that OLS regression often is robust to small to moderate violations of its assumptions. What we need to determine is whether the nature and amount of violation is consequential. Few tests of assumptions are amenable to doing so in an intuitive way. Third, many tests of assumptions are based on asymptotic theory and only perform adequately with large sample sizes (Shapiro & Wilk, 1965). However, with large N, preliminary significance tests can detect minor departures from assumptions that are of little consequence. You are in a no-win situation in that if your N is too small, the p values for the preliminary test will be incorrect; if your N is too large, you run the risk of inappropriately concluding the assumption violation is consequential when, in fact, it is not. Fourth, different tests of non-normality are sensitive only to certain forms of non-normality, i.e., some tests do a reasonably good job of detecting kurtosis but not skewness and for other tests, the reverse is the case. Similarly, tests of heterogeneous variances can be sensitive to only certain types of heterogeneity. Fifth, preliminary tests often make assumptions in their own right and may perform poorly when their assumptions are violated. For example, many tests of variance heterogeneity make normality assumptions and are not robust to violations of them. Sixth, using preliminary tests as a screen can change sampling distributions in unpredictable ways. For example, the statistical theory of t tests for comparing two independent means was derived without the idea of first applying a screening test for normality. Introducing this preliminary step no longer allows all possible random samples of a given size to be part of the sampling distribution. Instead, we are allowing only samples that have passed the screening test to be part of the sampling distribution. The new sampling distribution may no longer be distributed as t, but we still erroneously use the t distribution as the reference for calculating the p value and confidence interval.

A growing number of statisticians recommend against the two-step strategy that uses preliminary tests as screeners. The preferred approach is to use methods that do not make the assumptions of OLS to begin with, such as by using regression-based bootstrapping, Huber-White estimation (Keselman et al., 2013) or some other robust estimation approach. I adopt this latter strategy. Despite this, it is important to explore regression assumptions to understand the broader analytic scenario you are in. On my website, I provide a program called *regression diagnostics* that executes a range of graphical methods for exploring regression assumptions.

In addition to assumption violations, another concern with traditional regression methods is that they are outlier sensitive. An outlier is a data point that distorts basic trends in the data. The susceptibility of regression parameters to outliers is reflected in a concept called the **finite sample break down point**. The finite sample break down point is the minimal proportion of observations that, when altered sufficiently, render a statistic meaningless (Wilcox, 2017). As an example, if in a set of N scores, a single score equaled infinity, then the value of the mean of those scores would be infinity and meaningless. The finite sample breakdown point of the mean is 1/N because a single aberrant score out of the N scores can render it meaningless. This also is true of a standard deviation, a correlation, and a regression coefficient. A median, by contrast, has a finite sample break down point of 0.50. For example, the median of the scores 1, 3, 5, 7, and 9 is 5. The median of the scores -999999, 3, 5, 7, and 9 also is 5. No matter how extreme we make the scores around the midpoint, the median remains meaningful. A 20% trimmed mean, which is calculated by trimming away the top and bottom 20% of cases in the distribution, has a finite sample breakdown point of 0.20. Robust statistics seek to work with parameters with large finite sample break down points, i.e., outlier resistant parameters. I revisit outlier resistant regression methods in Chapter XX.

Random versus Fixed Predictor Regression

Although often not discussed in statistical textbooks, there actually are two different statistical models for multiple regression. The first model, called a **fixed predictor model**, treats predictors as having fixed values. By fixed, I mean that the researcher specifies the values of the predictors that are of interest and then seeks to make inferences about the outcome relative to those specific predictor values. For example, for the predictor variable of biological sex, a researcher decides to focus on the values of “male” and “female.” All values of interest on the predictor variable are in the dataset. The outcome variable in fixed predictor regression models is a random variable that is normally distributed at any given set of predictor values. The predictor values, however, are fixed or treated as such. Timm and Weisner (2003) refer to this model as the classic normal regression (CNR) model and it is the model I described above. Given its popularity, Sampson (1974) refers to it simply as “regression analysis.”

The second regression model treats both the outcome and predictors as being multivariate normally distributed in the broader population of interest. This is a stronger assumption than that of error score normality at a given set of predictor values per the CNR model. As well, the values of both the outcome and predictors are assumed to be randomly sampled from this larger population when individuals are randomly sampled for study. For example, annual income might be a predictor but my sample might include only a subset of income values from all possible values of income in the population I am studying. In this sense, the specific values of the predictors are not fixed; rather, they depend on who happens to be selected into the study. This model is often called a **random predictor model**. In contrast to the CNR model, interest is in making inferences about all values of the predictor that occur in the broader population, not just those values represented in the sample. The statistical theory for the model takes this into account and the theory is distinct from the CNR model. Timm and Weisner (2003) refer to this model as the jointly normal multiple linear regression (JNR) model. Sampson (1974) refers to it as “multivariate regression analysis.” The two models also are sometimes referred to as regression with nonstochastic (fixed) and stochastic (random) predictors, respectively.

In both models, there is a regression coefficient associated with each predictor and it is this regression coefficient that we are interested in estimating (in addition to an intercept). Most textbooks present the fixed predictor version of regression and most computer software uses fixed predictor statistical theory. Inferences and parameter estimates in the frameworks often yield identical results, so in this sense, the choice of the model does not matter, but some subtle differences can occur (Sampson, (1974). For example, Gatsonis and Sampson (1989) note that Cohen's (1988) approach to power analysis in multiple regression is based on a nonstochastic model that only approximates power estimates for multiple regression with random predictors, although the approximation is generally close. Kelley (2007) shows that the confidence interval for a squared multiple correlation can differ somewhat for the two model forms.

In practice, fixed predictor regression is most often used in the social sciences even when the predictors are quantitative and many valued. My focus will be on this model because of its popularity. For further discussion of the two models, see Sampson (1974) and Timm and Weisner (2003). I revisit the topic in later chapters.

Summary

In sum, traditional linear regression can be thought of in terms of the analysis of conditional outcome means and how those means vary across values of predictors. Linear regression assumes outcome means change as a linear function of predictor values. The coefficients in a regression analysis reflect how much the means are presumed to change given a one unit change in the predictor, holding other predictors in the equation constant. For quantitative predictors, for every one unit the predictor changes, the outcome mean is predicted to change by *b* units, where *b* is the estimated coefficient for the predictor. For dummy variables, one thinks of the value of the coefficient as the predicted mean for the group scored 1 on the variable minus the predicted mean of the reference group. Many researchers do not think of regression as analyzing outcome means, but it does so.

Traditional regression analysis makes population assumptions which sometimes are violated. Although it is commonly believed that OLS regression is robust to assumption violations, this often is not the case. When the assumptions are questionable and deemed consequential, robust regression methods should be considered. Judicious use of linear regression can yield estimates of (a) total effects of a program, (b) direct effects of a program on a mediator, (c) direct effects of a mediator on an outcome, (d), indirect effects of a program on an outcome through a given mediational chain. and (e) direct effects of a treatment on the outcome independent of the mediators. One can implement either fixed or random predictor regression models, with the former being more common.

THE BASICS OF BINARY REGRESSION

In SEM, binary regression is used when an outcome is dichotomous and the outcome is predicted from mediators, treatment dummy variables, and/or covariates. Binary regression also is used when a mediator is dichotomous and it is predicted from treatment dummy variables and covariates. The fundamental parameter of interest in binary regression is an outcome probability, such as the probability of contracting a disease, of being fired from a job, or of getting married. In binary regression, the probabilities are operationalized as the proportion of people in the population who have a score of Y = 1 for 0, 1 scoring of the outcome. An alternative way of expressing a probability is as an odds. We convert a probability to an odds by dividing it by 1 minus the probability in question. If the probability of 50 year old men in the United States seeing a doctor in the ensuing 12 months is 0.667, then the probability of not doing this is 1 - .667 = 0.333. The ratio of these two probabilities is the odds; 0.667/0.333 = 2.0, or in common parlance, the odds are "2 to 1" that 50-year-old men will see a doctor in the next 12 months.

An odds can be less than one. If the probability of a teenager smoking marijuana is 0.20, the odds of a teenager smoking marijuana is 0.20/0.80 = 0.25. This means the probability of smoking marijuana is one fourth the probability of not smoking marijuana. If the probability of a Black man being convicted of a crime is 0.25, the odds of a Black man being convicted of a crime is 0.25./0.75 = 0.33. The odds value of 0.33 means the probability of a Black man being convicted is one third that of not being convicted.

One can convert a probability to an odds by applying the formula P(Y) / (1-P(Y)), where P(Y) refers to the probability of Y. One can convert an odds to a probability by the formula Odds(Y) / (Odds(Y) + 1).

Binary Regression and the Modeling of Conditional Probabilities

Suppose we are interested in modeling how the probability of some event, Y, changes as a function of the values of one or more predictor variables, X. Consider the bivariate case where the outcome is the probability adolescents smoke marijuana in the ensuing year (Y) and the predictor is the age of adolescents (X), ranging from 12 to 17. We want to characterize what the probability of Y is for adolescents who are age 12, what it is for adolescents who are age 13, what it is for adolescents who are age 14, and so on. Each age represents a different predictor "profile" and we seek to characterize P(Y|X) at each profile, i.e., P(Y|X=12), P(Y|X= 13), P(Y|X=14), and so on.

One strategy for expressing how the probability of Y, in this case smoking marijuana, varies as a function of age, is to use a linear equation:

$$P(Y) = \alpha + \beta X \qquad [5.5]$$

where P(Y) is the probability that the specified group of adolescents smokes marijuana in the next 12 months and X is a variable reflecting the age groups of interest. Suppose the (population) probabilities of smoking marijuana are as follows:

Age	P(Y)
12	0.025
13	0.050
14	0.075
15	0.100
16	0.125
17	0.150

The probability of smoking marijuana is 0.025 conditional on age being "12." The probability of smoking marijuana is 0.050 conditional on age being "13." And so on. The intercept for this model is -0.275 and the slope is 0.025. Note that for every one unit age increases, the probability of smoking marijuana increases by 0.025 units. The intercept is meaningless in this case because it refers to an age (age = 0) outside the range of X.

In traditional regression analysis, the "error variance" in a linear equation is the variability of the Y scores at a given predictor profile. The variance of Y is assumed to be the same at each predictor profile per the assumption of variance homogeneity. In a linear probability model, the homogeneity of variance assumption is almost always violated. A formula for calculating the variance of dichotomous Y at a given predictor profile is P(Y)/(1-P(Y)). For example, if a profile has an outcome probability of 0.10 (i.e., a mean of 0.10 for the dichotomous outcome), the variance of the scores is 0.10(0.90) = 0.09. Given that the probabilities vary with age, the error variance also must vary with age. This lack of variance homogeneity complicates the calculation of standard errors, p values and confidence intervals. Parenthetically, note that in Equation 5.5 there is no disturbance/error term formally represented. This is because (a) the value of it is completely determined by the mean of Y and, hence, it is redundant and (b) the focus is on the prediction of the mean of Y, not with individual variability about that mean.

Because the relationship between age and the probability of smoking marijuana is linear, a potentially appropriate model for analyzing the data is the **linear probability model**, which assumes a linear function as reflected by Equation 5.5. An early strategy for analyzing the linear probability model (LPM) was to use OLS regression with a binary outcome. This approach capitalized on the fact that the mean of a dichotomous variable scored 0-1 equals the proportion of scores that have a 1. The OLS approach is problematic, however, because for standard errors and confidence intervals to be correct, OLS requires (1) the error scores for a predictor profile are normally distributed, which is not true for a binary outcome, and (2) the population error scores have equal variance across predictor profiles, which also is not the case.

It recently has been suggested that the LPM approach is workable in many cases if one uses OLS to estimate coefficients but instead of using traditional standard errors, one uses a robust estimator, such as a Huber-White estimator. This approach has been called the **modified linear probability model** (MLPM). A substantial literature supports the use of the MLPM for binary regression (Angrist & Pischke, 2009; Breen, Karlson, & Holm, 2018; Chatla & Shmueli, 2013; Cheung, 2007; Deke, 2014; Gomila, 2020; Hellevik, 2009; Horace & Oaxaca, 2006; Huang, 2022; Judkins & Porter, 2015; Pischke, 2012; Tsiatis et al., 2008; Uanhoro, Wang, & O'Connell, 2019; Von Hippel, 2015, 2017; Woolridge, 2010), although, like other binary regression models, there are scenarios where it fails (Chen, Martin & Woolridge, 2023; Lee, Lee & Choi, 2023). For example, one problem with the MLPM is that predicted scores using the equation can be (but are not always) outside the range of 0.0 to 1.0. When this occurs, the out-of-range cases are called **offending estimates**. Horrace and Oaxaca (2003, 2006) show that the OLS coefficients can be biased and inconsistent if there are too many such offending estimates. They recommend deleting such individuals from the analysis and then re-estimating coefficients and their standard errors using the MLPM. The idea is that if the predicted probabilities are construed as stand-ins for the true fitted values, then only the data with predictions between zero and one are needed to estimate coefficients. Horrace and Oaxaca refer to this approach as **sequential least squares** (SLS). It is supported by several simulations (Horrace & Oaxaca, 2003; Uanhoro et al., 2019). You will encounter disparaging remarks about the linear probability model, but such remarks typically apply to the unmodified OLS version of it. I address the criticisms in more detail below and elaborate more on MLPM perspectives. Keep in mind that the MLPM is potentially viable as long as the relationship between outcome probabilities and predictor values are approximately linear. If the functions are decidedly non-linear, then the use of the MLPM can lead to non-trivial estimation problems, especially when the outcome base rate is far from 0.50 (see Domingue et al., 2022).

Probably the most popular method for analyzing binary outcomes is **logistic regression**. It does not model probabilities, but rather odds. More technically, it models the log of odds. Let the odds of Y be represented by Odds(Y). Recall Odds(Y) is the P(Y) divided by 1 – P(Y). The logistic model relates the log of Odds(Y) to X as follows:

$$\ln[\text{Odds}(Y)] = \alpha + \beta X$$

where ln is the natural logarithm. This model posits that the log of the odds of Y is a linear function of X, whereas the MLPM states that it is the probability of Y, not the log odds of Y, that is a linear function of X. If the log of the outcome odds is a linear function of X, then, by definition, the probability of the outcome cannot be a linear function of X. Conversely, if the probability of the outcome is a linear function of X, then, by definition, the log odds of the outcome cannot be a linear function of X. By modeling the log odds as a linear function of X, the logistic model implies a non-linear relationship between X and the probability of Y. Consider our example on smoking marijuana but with log odds:

Age	P(Y)	Odds(Y)	ln(Odds(Y))
12	0.025	0.0256	-3.665
13	0.050	0.0526	-2.945
14	0.075	0.0811	-2.512
15	0.100	0.1111	-2.197
16	0.125	0.1429	-1.946
17	0.150	0.1765	-1.734

Note in this case that age is not linearly related to the log odds of Y. The logistic model is misspecified and not appropriate. In this sense, the logistic model is often called a non-linear model, at least with reference to the modeled probabilities.

It turns out that the above logit function, when translated into probabilities, implies a probability curve that is S shaped and takes a form known as a sigmoid function. Figure 5.2 shows an example of a probability curve for a logistic model across the majority of the probability spectrum as a function of a continuous X. Note there is a floor effect after which the probability of Y increases as a function of X. Eventually, a ceiling effect occurs, resulting in a flattening of the curve. It is not necessary for a continuous predictor to invoke the full range of probabilities in logistic regression. If the X invokes probabilities from 0.25 to 0.75, for example, the logistic function is basically linear. In this sense, logistic regression can accommodate linear relationships depending on the range of the probabilities in the data. Some scientists argue an S shaped function is more reasonable to expect for probabilities than linear functions. In my 40 years of working with a wide range of data, I have rarely encountered fully S shaped curves. Instead, I typically observe approximately linear curves or non-linear curves with a single bend representing either a floor or ceiling effect. In my opinion, we should not assume by fiat that a certain type of curve exists. The operative curve needs to be explored and a model chosen accordingly. For example, I find I can sometimes use a MLPM model with a quadratic term for X to capture a single bend (see Chapter 6).

Closely related to the logistic model is the **probit model**. Probit regression targets probabilities but models them by converting probabilities to values in a cumulative normal distribution. These Z values (rather than probabilities or log odds) are then modeled. The probit function also yields S shaped probability curves. The curve is similar to that of logistic regression, which is why some statisticians treat them as functionally interchangeable. However, there are differences as shown in Figure 5.2.[4] For details, see Long (1997); for other functional forms, see Long (1997), Dobson (2008), and Wooldridge (2010).

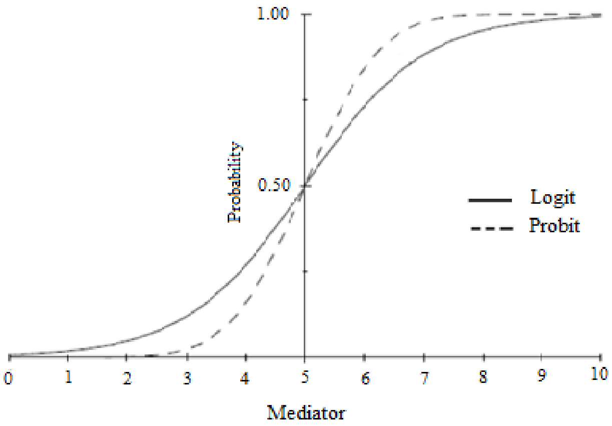

FIGURE 5.2. Probability functions for logistic and probit regression

Coefficients in Binary Regression

When we conduct a logistic or probit regression, we obtain the same output we would in a standard regression analysis, but instead of regression coefficients, we get logistic coefficients or probit coefficients. In this section, I describe how to interpret coefficients in binary regression.

The Modified Linear Probability Model

For the MLPM, coefficients are interpreted as in standard linear regression. Suppose I analyze whether parents obtain a vaccination for their child as a function of participating

[4] Just as the "normal distribution" refers to a family of distributions, this also is true of logit and probit distributions. As well, just as linear functions can vary in how flat or steep the line is when characterizing the relationship between two variables, such is also the case for the logit and probit functions.

in a program to educate parents about the benefits of the vaccination. There is a treatment group and a control group that does not participate in the program. The treatment group is represented by a dummy variable, D_T, with 1 = participated in the program and 0 = participated in the control group. The outcome variable, Y, is 0 = did not obtain a vaccination within six months, 1 = obtained a vaccination within six months.

Dummy Variables in the MLPM: Example of Total Effect Analysis. Suppose the MLPM yields the following results when I regress the binary outcome onto the dummy variable:

$$P(Y) = 0.51 + 0.25\ D_T \qquad [5.6]$$

The intercept in the above equation is the predicted mean outcome when all predictors = 0. When $D_T = 0$, the intercept is the control group mean. Since the mean Y is analogous to a probability (given Y is scored 0 or 1), the probability or proportion of parents in the control group who obtained the vaccination is 0.51. For dummy coded predictors, the coefficient in the MLPM is the estimated Y mean difference between the group scored 1 on the dummy variable and the reference group. Thus, the treatment group mean is 0.25 units higher than the control group mean, i.e., the proportion of parents in the treatment group who obtained the vaccination is 0.51 + 0.25 = 0.76.

Continuous Variables in the MLPM: Example Mediator Analysis. For continuous predictors, the MLPM coefficient reflects the predicted change in the outcome probability given a one unit increase in X, holding all other predictors in the equation constant. Suppose I measure for each parent an overall judgment of the perceived positives and negatives of obtaining the vaccination on a -3 to +3 scale where the metric is -3 = very negative, -2 = moderately negative, -1 = slightly negative, 0 = neutral, 1 = slightly positive, 2 = moderately positive, and 3 = very positive. This is a presumed mediator of the program effect and I refer to it as M. My goal is to document the estimated effect of the mediator on the outcome. Using the MLPM, I regress Y onto M and D_T and I obtain the following results:

$$P(Y) = 0.49 + 0.15\ M + .01\ D_T \qquad [5.7]$$

The coefficient for M indicates that for every one unit the perceived positives and negatives judgment increases, the probability or proportion of parents who obtain the vaccination for their child increases by 0.15, holding constant all other predictors in the equation, namely D_T. The intercept is the estimated proportion of parents who obtain the vaccination in the control group (because $D_T = 0$) given that their perceived positives and negatives score is 0. This is because the intercept is the mean outcome when all predictors equal zero. The coefficient for D_T is the outcome mean difference between the treatment and control group when the perceived positives and negatives mediator is held constant at a specific value (e.g., when parents have a score of +1 on M or when parents have a score of -1 on M, and so on). It differs in value from the coefficient for D_T in Equation 5.6 because in that equation, I did not hold the mediator constant. In this case, the coefficient for D_T reflects the effect of the treatment on the outcome over and above the effects of the mediator. The effect is minimal (there is a 1% differences).

Table 5.2 provides the calculated predicted probabilities for each value of M for the treatment and control groups. I calculated these values using Equation 5.7 and substituting the relevant values of M and D_T into it. For example, the predicted probability for people in the control group who have an M score of -3 is

$$P(Y) = 0.49 + 0.15\ (-3) + .01\ (0) = 0.0400$$

Table 5.2: Results of MLPM Analysis of Mediator Effect

Value of M	Control Probability	Treatment Probability
-3	.0400	.0500
-2	.1900	.2000
-1	.3400	.3500
0	.4900	.5000
1	.6400	.6500
2	.7900	.8000
3	.9400	.9500

Note that the probabilities in each group are a linear function of M; for every one unit M increases, the probability of or proportion of parents obtaining a vaccination for their child increases by 0.15 units, the value of the coefficient for M.

Logistic Regression

Dummy Variables in Logistic Regression: Example of Total Effect Analysis. For logistic regression, here is the equation that results when I analyze parents obtaining a vaccination for their child as a function of participating in the program to encourage vaccinations versus being in the control group:[5]

[5] I extend the results to four decimals to avoid rounding error for some of the points I want to make.

$$\ln[\text{Odds}(Y)] = 0.0400 + 1.1527\ D_T \qquad [5.8]$$

The intercept is the predicted log odds of parents obtaining the vaccination when all predictors equal 0. Because $D_T = 0$ designates the control group, the intercept is the predicted log odds of obtaining a vaccination for the control group. To make the intercept more meaningful, researchers often compute the exponent (or anti-log) of it, which yields exp(0.0400) = 1.0408. The odds that parents in the control group obtain a vaccination are just over "1 to 1." I can convert this odds to a probability by dividing it by the odds plus one. The probability or proportion of parents obtaining a vaccination for the control group is thus 1.0408/(1.0408+1) = 0.51, which agrees with the result from the MLPM in Equation 5.6. In applications of logistic regression, most researchers work with odds. I personally prefer to work with probabilities because I find them more intuitive. Given this, I often make conversions to probabilities even when using logistic modeling.

For reasons that will be apparent shortly, I also can use Equation 5.8 to calculate the log odds, odds, and probability of parents obtaining a vaccination if they participated in the treatment program. This group has a value of $D_T = 1$, which yields

$$\ln[\text{Odds}(Y)] = 0.0400 + 1.1527\ (1) = 1.1927$$

The exponent of 1.1927 is the odds of these parents obtaining a vaccination for their child, which equals exp(1.1927) = 3.2960; it is over three times more likely that parents exposed to the program will obtain the vaccination than not when likelihoods are framed as odds. Converting the odds to a probability yields 3.2960/(3.2960+1) = 0.76; 76% of the parents in the treatment condition obtained the vaccination for their child. This is the same result I found in the MLPM using Equation 5.6.

The logistic coefficient for D_T was 1.1527. This is the predicted log odds of obtaining a vaccination for the group scored 1 on the dummy variable minus the corresponding log odds for the reference group, namely 1.1927 - 0.0400. To make this coefficient more interpretable, researchers often calculate the exponent of it, which yields exp(1.1527) = 3.1667. This value is an **odds ratio** because it equals the odds of obtaining a vaccination for the group scored 1 on D_T *divided by* the odds of obtaining a vaccination for the reference group. Recall that the odds of obtaining a vaccination for the treatment group was 3.2960 and for the control group it was 1.0408. The ratio of these odds is 3.2960/1.0408 = 3.1667, which is the exponent of the coefficient for D_T. It is just over three times more likely that parents in the treatment group will obtain a vaccination for their child than parents in the control condition, when the likelihood is framed as odds. Note that if the probability of parents obtaining a vaccination is identical for the treatment and control groups, then the odds will be identical and the odds ratio will equal 1.0.

Whereas when comparing means in traditional regression a difference of zero implies no program effect, for odds ratios, a value of 1.0 implies no program effect.

In sum, the exponent of the logistic coefficient for a dummy variable is an odds ratio that divides the odds of engaging in the outcome for the group scored 1 on the dummy variable divided by the corresponding odds for the reference group. The odds comprising the numerator and denominator of the odds ratio can be calculated from the logistic equation (in this case, Equation 5.8) and these odds can be converted to probabilities/proportions, if desired.

Continuous Variables in Logistic Regression: Example Mediator Analysis. For a continuous predictor, the logit coefficient is the number of log odds that the outcome is predicted to change given a one unit increase in the predictor. For the equation using the mediator (perceived negative versus positive consequences) and treatment condition as predictors of parents obtaining a vaccination, the logit equation is:

$$\ln[\text{Odds}(Y)] = -0.0400 + 0.6152\ M + .0400\ D_T \qquad [5.9]$$

Using this equation, I want to document the effect of the mediator on the outcome. As before, the intercept is the predicted log odds of control parents obtaining a vaccination for their child (because $D_T = 0$) when M = 0. The exponent of the intercept, -0.0400, is exp(-0.0400) = 0.9608; it is the odds of control parents obtaining a vaccination. The odds are about “1 to 1,” or it is about as likely that the parents will obtain the vaccination as not. Converted to a probability, it is 0.9608/(0.9608+1) = 0.49, which agrees with the result from the MLPM in Equation 5.7; the proportion of parents who obtain a vaccination for their child in the control group and who have a score of M = 0 is 0.49.

The exponent of the coefficient for M is the **multiplicative factor** by which the predicted odds of Y changes given a one unit increase in M, holding the other predictors constant. For example, the exponent of 0.6152 is exp(0.6152) = 1.85. Recall that the control group odds of obtaining a vaccination when M = 0 was 0.9608. If M increases by 1 from 0 to 1, then the predicted odds of control parents obtaining a vaccination change from 0.9608 to (0.9608)(1.85) = 1.777. If M increases again by 1 unit (from 1 to 2), then the predicted odds of control parents obtaining a vaccination change to (1.777)(1.85) = 3.288. And so on. Table 5.3 provides the calculated predicted log odds, odds and probabilities for each value of M for the treatment and control groups using Equation 5.9.

The column labeled “Odds” is the exponent of the entries in the column “Log Odds” just to the left of it. The column labeled “Probability” equals the entries in the “Odds” column divided by the “Odds” entry plus 1. Note that (a) the “Log Odds” entries are a linear function of M, increasing by 0.6152 (the value of the coefficient for M) in each successive row; (b) a row entry in the “Odds” column is equal to the entry in the prior row times 1.85, the multiplicative factor identified above, and (c) the entries in the "Probabilities" columns are a non-linear function of the values of M.

Table 5.3: Results of Logistic Analysis for Mediator

Value of M	Control Log Odds	Control Odds	Control Probability	Treatment Log Odds	Treatment Odds	Treatment Probability
-3	-1.8856	.1517	.1317	-1.8456	.1579	.1364
-2	-1.2704	.2807	.2192	-1.2304	.2922	.2261
-1	-.6552	.5193	.3418	-.6152	.5405	.3509
0	-.0400	.9608	.4900	.0000	1.0000	.5000
1	.5752	1.7775	.6400	.6152	1.8500	.6491
2	1.1904	3.2884	.7668	1.2304	3.4226	.7739
3	1.8056	6.0836	.8588	1.8456	6.3319	.8636

In sum, the exponent of the coefficient for a continuous predictor is the multiplicative factor by which the odds change given a one unit increase in the predictor. If the multiplicative factor equals 1.00, the predicted odds do not change with increases in X because any number multiplied by 1.0 equals itself. If the multiplicative factor is greater than 1.0, the predicted odds increase with a one unit increase in the predictor. For example, a multiplicative factor of 2.0 means the odds double with each unit increase of the predictor; a multiplicative factor of 3.0 means the odds triple with each unit increase of the predictor. If the exponent of the coefficient is less than 1.0, then the predicted odds decrease with a one unit increase in the predictor. For example, if the multiplicative factor is 0.50, then the odds cut in half when the predictor increases by one unit; if the multiplicative factor is 0.33, then the odds cut by a third with a one-unit increase.

Probit Regression

For probit regression, one must keep in mind scores in a cumulative standard normal distribution when interpreting coefficients. For convenience, I refer to these scores as Z scores. They are indicative of the area under the curve in a cumulative standard normal distribution that is less than Z. In such a distribution, a Z score of -1.96 reflects the case where 0.025 (or 2.5%) of the scores are less than it and 0.975 (or 97.5%) of the scores are greater than it. A Z score of 0.0 is such that 0.50 of the scores are less than it and 0.50 of the scores are greater than it. A Z score of 1.0 is such that 0.66 of the scores are less than it and 0.34 of the scores are greater than it. And so on. On my website, you can translate a

Z score into a probability in a cumulative standard normal distribution using the p values program on the Program tab.

Dummy Variables in Probit Regression: Example of Total Effect Analysis. For a dummy variable with 0-1 coding, the coefficient in a probit regression is the difference between the predicted Z score for the group scored 1 on the dummy variable minus the predicted Z score for the reference group, holding constant the other predictors in the equation. Here is the equation from the parent vaccination example predicting vaccination behavior from the treatment versus control condition:

$$\text{Probit}(Y) = 0.025 + 0.680\ D_T \qquad [5.10]$$

Where I use the term "Probit" to refer to a transformation to a Z score in a cumulative normal distribution (CDF). The intercept is the Z score in a CDF for the control group parents. I can covert this into a probability by determining the proportion of cases below a Z score of 0.025 using the p value program on my website, which is 0.51. This is the proportion of parents in the control group who obtained the vaccination for their child and agrees with both the MLPM and the logistic model. The predicted Z score for the treatment group is

$$\text{Probit}(Y) = 0.025 + 0.680\ (1) = 0.705$$

This Z score of 0.705 converts to a probability of 0.76, which also agrees with the MLPM and logistic models. The effect of the treatment in probability metrics is the probability for the treatment group minus the probability for the control group and is 0.76 – 0.51 = 0.25. Thus, I can express the effect of the program in terms of differences in Z score units (0.680) or in probability units (0.25).

Continuous Variables in Probit Regression: Example Mediator Analysis. For a continuous predictor, the probit coefficient is the number of Z scores that the outcome is predicted to change given a one unit increase in the continuous predictor holding other variables in the equation constant. For the equation using the mediator and treatment condition as predictors of parents obtaining a vaccination, the probit equation is:

$$\text{Probit}(Y) = -0.025 + 0.384\ M + 0.025\ D_T \qquad [5.11]$$

My goal, again, is to document the estimated effect of the mediator on the outcome net the effect of the treatment condition. The intercept is the predicted Z score of parents obtaining a vaccination for their child in the control group (because $D_T = 0$) when $M = 0$. The Z score of -0.025 translates in a cumulative standard normal distribution into a probability of 0.49, which agrees with the results of the MLPM and logistic regression.

The coefficient for M is the number of Z scores that the outcome is predicted to change given a one unit increase in M, holding the other predictors constant. If M increases by 1 from 0 to 1, then the predicted Z score for control parents obtaining a vaccination changes from -0.025 to 0.359 (because -0.025 + 0.384 = 0.359). This translates into a probability of 0.640 a difference of 0.15 probability units from the case where M = 0. Table 5.4 provides the predicted Z scores and their probabilities for each value of M for the treatment and control groups based on Equation 5.11:

Table 5.4: Results of Probit Analysis for Mediator

Value of M	Control Z Value	Control Probability	Treatment Z Value	Treatment Probability
-3	-1.177	.120	-1.152	.125
-2	-.793	.214	-.768	.221
-1	-.409	.341	-.384	.350
0	-.025	.490	.000	.500
1	.359	.640	.384	.650
2	.743	.771	.768	.779
3	1.127	.870	1.152	.875

The entries in the column labeled "Z Value" are a linear function of M, increasing by 0.384 in each successive row. The entries in the "Probabilities" column are a non-linear function of the values of M. This is the non-linearity inherent in probit regression.

In sum, when analyzing binary outcomes, three commonly applied approaches are the MLPM, logistic regression, and probit regression. The interpretation of coefficients in each framework is different but their results often converge. I describe in Chapter XX how to decide which approach to use in RETs.

Average Marginal Effects

For logistic and probit regression, some researchers report what are called **average marginal effects** (AME) for each predictor because of dissatisfaction with the properties and interpretation of odds ratios or Z scores (I elaborate these limitations in Chapter 12). Instead of focusing on odds or Z scores, AMEs describe how outcome probabilities vary as a function of predictors but the logic is still grounded in logit/probit regression. The values of AMEs often are similar to the values for coefficients yielded by the MLPM.

I can give you an intuitive sense of an AME using an example in which a binary mediator, M, maternal use of guilt as a discipline strategy (0 = no, 1 = yes), is a predictor of a binary outcome, Y, child depression (0 = not depressed, 1 = depressed). The hypothesis is that parents who rely on guilt induction are more likely to have a depressed child than parents who do not rely on guilt induction. Suppose I conduct a logistic regression predicting Y from M and a host of covariates that are included in the equation for control purposes. After conducting this analysis, I take the first case in the data set and I treat that child as having a mother who uses guilt as a discipline strategy irrespective of whether this truly is the case, i.e., I set the child's score equal to 1 on the guilt predictor. The child's scores on all the other predictor variables in the equation are left alone and reflect their naturally occurring values. I then calculate the predicted probability of depression for that child based on that child's scores on all of the predictors *using the initially derived logistic equation*. Specifically, I compute the predicted log odds of being depressed for the child but with the discipline strategy dummy variable set to one. I calculate the exponent of the predicted log odds to derive the predicted odds, and then I divide this odds by the odds +1 in order to convert it to a probability. Next, I repeat this process, but this time I presume the child has a mother who *does not* use guilt as a discipline strategy irrespective of whether that is actually the case, i.e., I set the score on the dummy variable for use of guilt to 0. I calculate the predicted probability of the outcome for the child under this scenario using the same logic as before. The difference between the two calculated probabilities is the marginal effect of the parental discipline strategy on child depression for that particular child. I repeat this process for every case/child in the sample, calculating a marginal effect for each one. Finally, I compute the average of all these individualized marginal effects. The result is a sample estimate of the population average marginal effect, AME, for the guilt predictor.

In essence, I am comparing two populations in the above process, one population in which every parent uses guilt as a discipline strategy and one in which every parent does not use guilt as a discipline strategy, but where each population has the same distribution of values on the other predictors. Because the only difference between the two populations is their value on the predictor of use of guilt, use of guilt must be the source of the differences in their likelihood of depression. Hence, this is the AME for use of guilt. For example, I might find that the AME for use of guilt as a discipline strategy is 0.30. This means that the proportion of children who report being depressed is 30% higher if their mother uses guilt as a discipline strategy than if their mothers do not.

If the predictor variable is continuous, an AME is similar to the above but it is calculated to reflect how the outcome probability changes if we increased everyone's score on the target predictor by one unit. For example, if the AME is 0.10 for a -2 to +2 scale that reflects how controlling the mother is, this means that for every one-unit controllingness increases, the probability of (or proportion of) children who are depressed increases by 0.10.[6] I delve into the technicalities of AMEs in Chapter 12; for now, I only seek to introduce the concept and provide an intuitive sense of it. For statistical details, see Wooldridge (2010). Again, some researchers highlight AMEs when interpreting the results of a logistic regression whereas others prefer to highlight odds ratios. For reasons I discuss in Chapter 12, I tend to prefer AMEs and conditional probabilities. I provide a program for computing AMEs in binary regression on my website.

The Latent Response Representation of Logistic Regression

Logistic and probit regression have been conceptualized in different ways, one of which uses statistical theory based on the generalized linear model (GLM). I used variants of that theory above to introduce the approaches. An alternative statistical theory has been used as a basis for logit and probit regression and I will make use of this theory in future chapters, in addition to the GLM approach. As such, I briefly characterize it here.

Consider the case where the outcome variable is dichotomous and focuses on a behavior we seek to have people perform (e.g., obtain a vaccination). The post-treatment outcome is scored 1 = performed the behavior versus 0 = did not perform the behavior. In the latent response approach, there is said to exist an underlying, unmeasured latent propensity, that I will signify as y*, that is continuous and reflects the propensity to engage or not engage in the dichotomous behavior.[7] When an individual crosses a threshold value on y*, represented by the parameter τ, his or her value on the observed dichotomous variable, y, changes from 0 (non-performance) to 1 (performance), i.e., s/he performs the behavior. For example, y* might range from -2 to +2 and the threshold value might be 0; if a person's location on y* is below 0, then the person does not perform the behavior. If the person's location on y* is zero or above, the person performs the behavior. Of theoretical interest is what the value of τ is for y* because it differentiates behavioral performance from behavioral non-performance.

In theory, we can construct causal models of the determinants of y* and, with concomitant knowledge of τ, we can then build an understanding of the dichotomous outcome. Like multiple regression, given a predictor set X, y* is assumed to be a linear function of the X:

$$y^* = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k + \varepsilon \qquad [5.12]$$

[6] Technically, AMEs for continuous variables focus on instantaneous change, a concept I introduce in Chapter 6.

[7] Do not confuse my use of the term "latent" here with the concept of a latent variable in structural equation modeling (SEM). Although there are similarities, the literature and meanings surrounding the term "latent" is distinct for logistic/probit regression compared to SEM.

In this equation, I use γ to represent a regression coefficient in place of the usual symbol β to signify that I am working with the latent response formulation. γ_0 represents an intercept. The last term in the equation is a random disturbance term that, like most regression models, is assumed to be independent of the X and has a mean of zero.

Theoretically, I am interested in estimating the various γ in Equation 5.12 because they provide insight into the causal impact of X on the propensity to engage in Y. When we conduct a traditional logistic regression analysis, we apply the model

$$\ln[\text{Odds}(Y)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad [5.13]$$

That is, we estimate the logistic coefficients rather than the γs in Equation 5.12. We ultimately want to link the various β to the γ so we can understand and estimate the values of γ. As well, we want to identify the value of τ. To do so is much easier said than done because we usually do not have enough information to accomplish it all. To do so, we have to make simplifying assumptions, many of which are somewhat concocted. I discuss the assumptions in the Appendix to this chapter. Appreciating the logic described in the Appendix will be useful for my discussion of the modeling of binary outcomes using SEM in future chapters, so I encourage you to work through it.

The Modified Linear Probability Model Revisited

Because I make use of it in future chapters and because it is somewhat controversial, I offer here a few more comments about the MLPM. The primary arguments in favor of using the MLPM include (1) it is computationally easy, (2) the coefficients in the model are subject to straightforward interpretation, (3) it often is easier to work with in complex mediational models than relying on the non-linear models of logit and probit regression, and (4) the patterns of statistical significance and the regression coefficients usually (but not always) are close in value to the average marginal effects calculated in logistic and probit regression.

The arguments typically leveled against the MLPM are several. First, as noted, some scientists argue that an S shaped function is more reasonable to expect for probabilities than linear functions. However, it is rare for such assertions to be substantiated with empirical evidence. As noted, I rarely encounter fully S shaped functions when working with probabilities. Obviously, MLPMs will not capture effects well if the true function in the population is decidedly non-linear (Domingue et al., 2022) but as Pischke (2012) notes, neither will a “wrong” non-linear model, such as when the observed non-linearity does not conform to that of a logit or probit function. To quote Pischke, “The fact that we have a probit, a logit, and the LPM [linear probability model] is just a statement to the fact that we don't know what the "right" model is. Hence, there is alot to be said for sticking to a linear regression function as compared to a fairly arbitrary choice of a non-linear one. Nonlinearity per se is a red herring." In the final analysis, if you expect in your data curvature in the probabilities of your outcome as a function of your predictors and that curvature is important to capture, then the MLPM might have to be altered to include power polynomials or some other non-linear modification to the model to respect and detect the non-linearity. Or, shift to logistic or probit regression if it better captures the curvature.

Of course, probabilities can't linearly increase forever as a function of increases in a continuous predictor X and eventually the probabilities will hit a ceiling of 1.0 at extreme values of X. In this sense, the MLPM cannot ever be said to represent the data generating function for truly continuous X. However, one can argue that *within the range of X values studied* and that occur in the population of interest, linear changes in probabilities as a function of X can indeed capture the causal dynamics at hand.

A second objection to the MLPM is that predicted values for the model can exceed 1.00 and be less than 0.00, which are nonsensical values for probabilities. This property, the argument goes, automatically eliminates the MLPM as a viable characterization of the data generating process linking the probability of Y and a continuous X. Critics who make this point, however, do not object to the fact that predicted values also occur outside the range of possible scores in standard OLS regression as applied to quantitative outcomes. For example, Y may have a metric with lower and upper bounds of 1 to 10, but predicted scores can occur in OLS regression that are less than 1 and greater than 10. Is there a double standard here? Values outside the probability metric of 0 to 1.0 are mostly a problem if one seeks to use the derived probabilities in prediction contexts or if they occur with sufficient frequency that estimation is undermined (Horace & Oaxaca, 2003, 2006). In a purely prediction context, if a person has a negative probability, one must decide what to do when making a prediction about behavioral performance or the outcome state for the individual. One logical action is to treat this probability as zero; if the probability is greater than 1.00, treat it as 1.00. For other interesting approaches to this problem, see Allison (2020). For non-prediction contexts, the occurrence of such scores, if problematic, often can be addressed by the methods of Horace and Oaxaca (2003, 2006) described above. As Wooldridge (2002) states in his classic text (p. 455), "If the main purpose is to estimate the partial effect of [the independent variable] on the response probability, averaged across the distribution of [the independent variable], then the fact that some predicted values are outside the unit interval may not be very important." In my own research across a wide range of data sets I have used for the past

40 years, I typically find high correlations between predicted probabilities based on logit or probit regression and those based on the MLPM, usually > 0.99.

A third objection to the linear probability model is that the disturbance term inherently has non-constant variance (heteroscedasticity) and is non-normal. The use of robust standard errors in the MLPM tends to mitigate this objection, especially with larger sample sizes.

Finally, some statisticians object that the MLPM can produce biased and inconsistent estimates and that the model can be undermined by measurement error in the outcome. These properties are matters of degree. Bias, inconsistency, and measurement error are not all or none. They vary in degree and the degree of bias or inconsistency that occurs may not be consequential in your particular analytic scenario.

Statistical purists object to the MLPM because it is theoretically inelegant and, in their view, a "wrong" model. Those who use the MLPM do not disagree with such characterizations but argue that the MLPM often can get the job done just as well as logit or probit regression, and sometimes better. To paraphrase Farnam (2011): "*Is it ever appropriate to forgo estimating the 'perfect' econometric model in order to gain something else on some other front? In economic theory, for instance, the 'best' models (i.e., those most used and treasured) are often not strictly correct in the sense that parsimony is valued. So, we throw out all types of real phenomena in order to get something simple, tractable, and easy to work with. That doesn't mean we don't tailor its sophistication and applicability depending on context, but we keep some simplification for the sake of clarity*." In the final analysis, researchers can engage in the machinations of logistic/probit regression for mediation and moderation analysis. However, the MLPM will often result in the same conclusions but without the headaches (see Chapter 12).

Chen, Martin and Woolridge (2023) emphasize the utility of average marginal effects for substantive research and note that the MLPM often is quite effective in approximating them, even when a substantial number of cases yield predicted probabilities outside the values of 0 and 1.00. In a series of simulations, Chen et al. (2023) found that the MLPM captured AMEs well when the explanatory variables were multivariately normally distributed. Estimation issues arose, however, for some types of asymmetric predictor distributions in which Horace-Oaxaca offending estimates crept into the analysis with sufficient frequency. Holm, Ejrnæs and Karlson (2015) note cases where the accuracy of the MLPM is affected by the distributional shape of the continuous predictors as well as truncation of the latent propensity underlying the binary outcome and model scale parameters (see also Lee, Lee, and Choi, 2023). Like other statistical methods, the performance of the MLPM is built on assumptions that may or may not be tenable in a particular research setting. Recognition of such limitations is important for all of our statistical approaches, not kust the MLPM.

The Log Binomial Model

A final binary regression model worth mentioning is the **log binomial model** (also called a **relative risk model**). It yields probability curves of the form shown in Figure 5.3. There is a floor effect followed by a rapid increase in the outcome probability. Log binomial models specify the log of the probability of Y as a linear function of X, which contrasts with the logistic model that specifies the log of the *odds* of Y as a linear function of X, namely ln[P(Y)] = α + β X compared to ln[Odds(Y)] = α + β X.

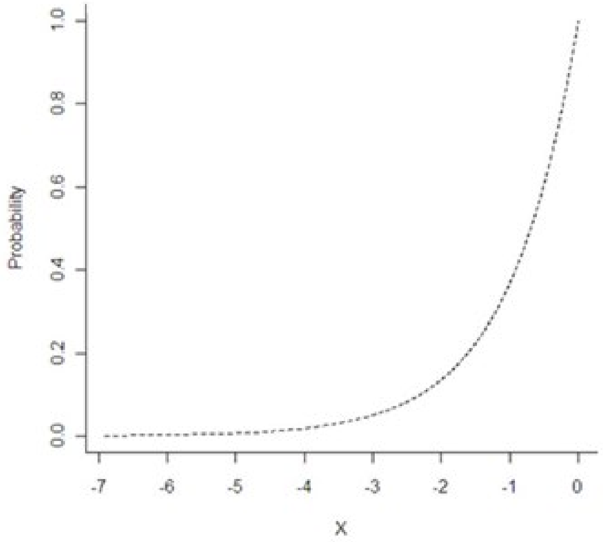

FIGURE 5.3. Log binomial model

Log binomial regression is used to model relative risks, an index I discuss in Chapter X. A complication of log binomial regression is that the *p* values for coefficients of a predictor can change when the outcome variable is reverse coded, from 0-1 to 1-0. Algorithms for the log binomial model also often fail to converge, especially with continuous predictors. It generally is recommended that one instead use a model that operates in the spirit of the log binomial model but that more readily converges, called **Zou's (2004) modified Poisson method** with robust Huber-White standard errors.

Binary Regression and the Generalized Linear Model

When reading articles about the above models, you may encounter references to **link functions**. This terminology comes from the statistical theory of **generalized linear models** (GLMs), which is not to be confused with what is often called the **general linear model**. I do not elaborate the theory here. I merely note you will encounter references to the above models in terms of their link functions in generalized linear models, e.g., a logit link function, a probit link function, an identity link function or a log binomial link function. For details of the theory, see Long (1997) and Wooldridge (2010).

In sum, there are a host of binary regression models in the statistical toolbox you can use to analyze dichotomous mediators and dichotomous outcomes in the context of RETs. Each has strengths and weaknesses. I discuss in future chapters some of the criteria you should use to choose between the different modeling strategies.

THE BASICS OF ORDINAL REGRESSION

Ordinal regression is applied in cases where the outcome variable is measured at an ordinal level. Scientists who design RETs for program evaluation exert influence on the measures they use and should seek to avoid crude forms of ordinal measurement. There is an extensive psychometric literature that provides guidance on how to construct interval-level measures (see Jaccard & Jacoby, 2020). However, there may be times when circumstances or conventions in the field demand the use of ordinal measures.

Oversimplifying somewhat, ordinal regression conducts a series of binary regressions between pairs of categories of the outcome variable. In general, if the outcome has *k* categories, ordinal regression will generate *k*-1 pairs of binary regressions for purposes of predicting and understanding the outcome. Ordinal regression methods differ in how the pairs are defined and the assumptions made about coefficient equivalence across the regressions. I develop the above ideas using a popular ordinal regression model called the **proportional odds models**. For other variants, see Agresti (2010) and Chapter 13.

My example focuses on an outcome that asks individuals to respond to the item "I feel I am capable of dealing with stress in my life." Responses are made on a four point disagree-agree scale with response categories of 1 = strongly disagree, 2 = moderately disagree, 3 = moderately agree, and 4 = strongly agree. The outcome is predicted from D_T, a dummy variable representing whether individuals are randomly assigned to a control (0) or treatment group (1) to reduce the negative effects of stress in their lives. It is expected that people in the program relative to the control condition will be more apt to indicate they are capable of dealing with stress as measured at the posttest.

As noted, when conducting ordinal regression, we focus on *k*-1 pairs of categories of the outcome. In the current example, there are four categories, so I analyze 4-1 = 3 pairs. The pairs in the proportional odds model are defined by combining categories on the outcome metric. Any given pair is defined as (a) a given target category and all categories above it versus (b) all categories below the target category. In our example, one "pair" is category 4 versus categories 1, 2, and 3 combined. Another pair is categories 3 and 4 combined versus categories 1 and 2 combined. The final pair is categories 2, 3, and 4 combined versus category 1. There is a logic behind these pairings, the key to which is to think about them in terms of "break points," which I now elaborate.

One theoretically interesting "break point" on the outcome metric is for the collapsed categories 3 and 4, which represent explicit agreement with the statement characterizing the outcome ("moderately agree" and "strongly agree") versus the collapsed categories of 1 and 2, which represent explicit disagreement with the statement (responses of "moderately disagree," and "strongly disagree"). Suppose I find that individuals in the treatment program are twice as likely to be in the "3 or 4" categories as compared to individuals in the control group. This means that individuals in the treatment group are twice as likely to endorse the statement about being able to deal with stress in their lives. When logistic regression is used to analyze this "pair" of categories, the exponent of the logistic coefficient associated with the treatment dummy variable will equal 2.0.[8] It is indicative of a positive effect of the program.

Suppose instead of the above break point, I examine a different one. I might define a break point as category 4 (strongly agreeing with the item) versus categories "1, 2, and 3" (not strongly agreeing with it). Suppose I again find that individuals in the treatment program are twice as likely to be in the "strongly agree" category than individuals in the control group, i.e., the exponent of the logistic coefficient is again 2.0 even though I focus on a different "break point." This also is consistent with a positive program effect.

Finally, consider the break point for categories 2, 3, 4 (not strongly disagreeing with the statement) versus category 1 (strongly disagreeing with the statement). I might find that yet again, individuals in the treatment program are twice as likely to be in the "2, 3, and 4" category than individuals in the control group, yielding a logit coefficient exponent of 2.0. Once again, this reflects a positive program impact.

Each of these "contrasts" are of substantive interest and an ordinal regression analysis provides perspectives on each of them by conducting separate logistic (or probit) regressions defined by them. An assumption of the proportional odds model is that the coefficient for a given predictor (e.g., the treatment versus control dummy variable) will not change irrespective of which breakpoint is being analyzed. In the above example, the logistic coefficient for D_T is the same in all three breakpoint analyses and it would equal

[8] Technically, when I use the phrase "twice as likely" I am referring to odds rather than probabilities.

0.693, the exponent of which is 2.0. The assumption of equal coefficients is known as the **parallel coefficient assumption** or the **proportionality assumption**.[9] Note that given the assumption, it is only necessary to report a single coefficient value for a predictor. Once you know it, you know the value of all the coefficients for the predictor for each breakpoint pair. As an example, the ordinal regression of the current example might be reported in the form of the following three logistic equations (where the categories to the left of "versus" are scored 1 and the categories to the right of "versus" are scored 0):

4 versus 1, 2, 3: $-2.197 + 0.693\ T$ [5.14]

3 and 4 versus 1 and 2: $-1.099 + 0.693\ T$ [5.15]

2, 3 and 4 versus 1: $0.405 + 0.693\ T$ [5.16]

Note that the exponent of 0.693 is 2.0, which indicates the odds of having a score of 1 on the outcome referenced by the equation is twice as large in the treatment group as the control group. Although the three equations have the same logistic coefficient for each predictor, there is a different intercept for each one. This allows the proportion of control respondents who score above the break point to vary depending on the pair in question. In Equation 5.14, the log odds of being in category 4 versus categories 1, 2 or 3 for the control group was -2.197. This translates into an odds of being in category 4 for control individuals of 0.11 and a proportion of 0.11 / (0.11+ 1) = 0.10. By simple subtraction, the proportion of control individuals who were in categories 1, 2, and 3 is 1 – 0.10 = 0.90. The log odds of being in category 4 for the treatment group in Equation 5.14 is -2.197 + 0.693(1) = -1.504. This translates into an odds of 0.22 and a proportion of 0.18. Table 5.5 summarizes the log odds, odds, and probabilities of being above the break point as opposed to below it for each equation and each treatment condition. Note in Table 5.5 that the treatment odds is twice the size of the control odds in each equation.

Table 5.5: Results of Logistic Analysis for Mediator

Contrast	Control Log Odds	Control Odds	Control Probability	Treatment Log Odds	Treatment Odds	Treatment Probability
4 vs. 3, 2, 1	-2.197	0.111	0.100	-1.504	0.222	0.182
4, 3 vs. 2, 1	-1.099	0.333	0.250	-0.406	0.666	0.400
4, 3, 2 vs. 1	0.405	1.500	0.600	1.098	3.000	0.750

[9] I describe ways of testing this assumption in Chapter 13.

Interestingly, just as one can express logistic or probit regression in terms of an underlying latent propensity, we also can do so for an ordinal outcome. Assume there is a continuous latent variable, y*, that reflects the outcome variable. In our example, it is the continuous construct of degree of agreement with the statement "I feel I am capable of dealing with stress in my life." The four-point response scale reflects this continuous construct, but the measure is a crude representation of it. In principle, one can specify a set of rules by which people's location on y* translates into responses on the four point rating scale. Analogous to logistic regression, I formulate a rule that if a person's score on y* is below a certain threshold value, that I call τ_1, then the response made on the rating scale by the person will fall into category 1, "strongly disagree." If a person's score on y* is above that threshold, but below a second threshold, τ_2, then the response made on the rating scale will fall into category 2, "moderately disagree." If a person's score on y* is above threshold τ_2, but below a third threshold, τ_3, then the response made on the rating scale will fall into category 3, "moderately agree." Finally, if the person's score on y* is above threshold τ_3, the response on the rating scale will fall into category 4, "strongly agree." If the outcome measure has *k* levels, there are *k*-1 thresholds. Figure 5.4 presents the dynamic graphically, showing the location of 4 different people on the underlying y* dimension and the response they would make (using a dashed arrow) on the rating scale given that location and the operative thresholds.

Like logistic regression, it is possible to specify a regression equation that expresses y* as a linear function of a set of predictors, X. In the stress treatment program example, the equation is:

$$y^* = \gamma_0 + \gamma_1 D_T + \varepsilon$$

We are interested in estimating the coefficient γ_1. This can be accomplished mathematically from the observed data by using the logic described in the Appendix but extended to multiple logistic equations. The coefficients in the y* equation are interpreted like any regression coefficient, namely, for every one unit that X increases, the coefficient is the number of units that y* is predicted to change, holding all other predictors constant. In the case of a dummy variable, the coefficient is the mean y* for the group scored one minus the reference group y* mean, holding all other predictors constant. I illustrate this approach as applied to RETs in Chapter XX.

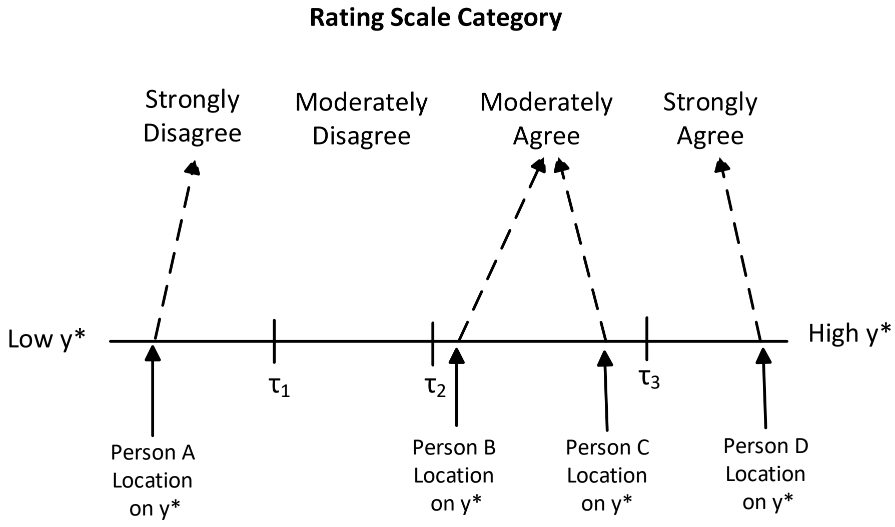

FIGURE 5.4. Illustration of thresholds

In sum, the primary ordinal regression analysis of RETs used in this book is grounded in logistic or probit regression but with breakpoints defined by the proportional odds model. I will invoke both an observed metric and latent response framework.

THE BASICS OF MULTINOMIAL REGRESSION

When an outcome variable is categorical and has more than two levels, analysis of it is typically undertaken using logistic-based multinomial regression. Consider the case where there are three categories on the outcome variable (Y=1, Y=2, and Y=3) and a continuous mediator predicting it, M. It is possible to conduct three logistic regressions based on all possible pairs of categories defined by the outcome variable. Specifically, I can predict Y=1 versus Y=2 from M, Y=1 versus Y=3 from M, and Y=2 versus Y=3 from M. When considered as a collective, the three equations are not independent and one would expect certain regularities across the equations. For example, knowing how M affects the log odds of Y=1 versus Y=2, as well as how M affects the log odds of Y=1 versus Y=3, tells us information about how M affects the log odds of Y=2 versus Y=3 (see Long, 1997, for elaboration). The multinomial regression model yields simultaneous estimates for the pairwise equations taking such dependencies into account. The "pairs" of outcome categories compared in the multinomial model take different forms, which the analyst specifies a priori. The output of the analysis is a series of logistic regressions representing each paired comparison. Each are interpreted as a traditional logistic model.

The most common multinomial model is called the **baseline-category model**. Given *k* levels of an outcome variable, one of the levels is declared by the researcher as the "baseline" or reference group. The analysis then estimates *k*-1 equations where each equation is a logistic regression model comparing the other levels of the outcome variable, separately, with the reference group. For example, suppose a developmental psychologist studies three types of attachment patterns that young children in dysfunctional families show with respect to their caretaker. The first type is secure attachment (SE) in which the child has a healthy, positive attachment to the caretaker. The second type is clinging (CL), in which the child excessively clings to the caretaker and shows patterns of unhealthy dependency. The third type is avoidance (AV), in which the child maintains distance from the caretaker and shows aloofness to him or her. Suppose a parenting program is developed to foster secure attachment. It does so by making the physical environment in the child's home more positive and by encouraging appropriate affectionate parenting on the part of the mother. The designers of the program hypothesized that higher levels of appropriate affection and a positive home environment would lead to higher odds of secure attachment relative to each of the other two forms of attachment. In my multinomial analysis I might declare secure attachment as the reference group. The multinomial model then performs two logistic regressions, one comparing AV versus SE categories and the second comparing CL versus SE categories.

There is an important property of multinomial logistic regression that is often overlooked. Consider the contingency table for the data in Table 5.6 that presents the number of children in each attachment category in the treatment and control conditions, respectively. There are a total of 300 children who received the program and 300 children who did not. The first two columns of Table 5.6 are the number of children who were in each attachment category at posttest as a function of treatment condition. The second two columns are the proportion of children in each category conditional on the treatment condition they participated in. I obtained these values by dividing the raw numbers in columns 1 and 2 by 300. For example, 50/300 0.167. This is the probability of exhibiting avoidance attachment given participation in the treatment program. The proportion 150/300 = 0.500 is the probability of exhibiting avoidance attachment given participation in the control condition. The fifth column in the table is the difference between the two proportions/probabilities. For example, the probability of or the proportion of children in the secure attachment category is 150/300 = 0.500 for children who participated in the treatment program but it is only 50/300 = 0.167 for children in the control group, a difference of 0.500 – 0.167 = 0.333. It is the probabilities in the first three columns that I am most interested in. I refer to them as **multinomial conditional probabilities**. Multinomial conditional probabilities, of course, have odds counterparts, which are simply the probability divided by one minus the probability. I also show these in Table 5.6 as well as the corresponding odds ratios for them as a function of treatment condition.

Table 5.6: Multinomial Conditional Probabilities

	Treat Freq	Cntrl Freq	Treat Prob	Cntrl Prob	Prob Diff	Treat Odds	Cntrl Odds	Odds Ratio
Avoidance	50	150	.167	.500	-0.333	.200	1.000	.200
Clinging	100	100	.333	.333	0.000	.500	.500	1.000
Secure	150	50	.500	.167	0.333	1.000	.200	5.000

Interestingly, the two pairwise logistic regressions yielded by a traditional multinomial regression analysis do *not* focus on the multinomial conditional odds nor the multinomial conditional probabilities. Rather, they focus on what are known as **local conditional odds** and **local conditional probabilities**. Consider the binary logistic regression model in a traditional multinomial regression that compares the categories of avoidance (scored 1) versus secure (scored 0). This analysis essentially ignores children who were clingers and analyzes only those children who are either in the avoidance or secure categories. Table 5.7 presents the Table 5.6 counterpart for this scenario. In this case, I calculated the probabilities for the treatment group by dividing the frequency by 200, namely the number of children in the treatment condition considering only avoidance and secure attachment children. I did the same for the control probabilities using the number of children in the control condition for the two groups summed as the divisor. Note that the sample sizes are different from those in Table 5.6 as are the estimated probabilities and the relevant probability differences. Table 5.8 presents the localized probabilities when the focus is on the clinging and secure children.

Table 5.7: Probabilities for Avoidance and Secure Children Only

	Treatment Number	Control Number	Treatment Probability	Control Probability	Probability Difference
Avoidance	50	150	.250	.750	-0.500
Secure	150	50	.750	.250	0.500

Table 5.8: Probabilities for Clinging and Secure Children Only

	Treatment Number	Control Number	Treatment Probability	Control Probability	Probability Difference
Clinging	100	100	.400	.667	-0.267
Secure	150	50	.600	.333	0.267

The question becomes which set of probabilities and contrasts should you focus your analysis on, the localized conditional odds and probabilities of Tables 5.7 and 5.8 or the multinomial conditional probabilities of Table 5.5? The answer is that it depends on your substantive foci and the questions you seek to address. For RETs, I usually focus on the multinomial conditional and probabilities and odds per Table 5.6. I show in Chapter 13 how to accomplish such analyses in an RET and discuss the advantages of the approach there. For more on the analysis of multicategory outcome variables in general, see Long (1997) and Agresti (1996).

In sum, when a mediator or outcome consists of a nominal variable with more than two levels, you typically will invoke multinomial logistic regression to model it. Within such modeling, you can focus on multinomial conditional probabilities, localized conditional probabilities, or both.

THE BASICS OF DISCRETE/COUNT REGRESSION

Sometimes outcomes we analyze are counts, such as the number of times in a year that a person has a doctor visit, the number of cigarettes someone smokes per day, and the number of times someone watches a given television series in the past month. Count outcomes are bounded at the lower end by zero and often they are non-normally distributed with substantial positive skew. For an outcome like the number of times young adolescents have used marijuana, the distribution usually has many observations at the value of zero because most young adolescents do not smoke marijuana, fewer observations at the value of one (some young adolescents have tried marijuana once), still fewer observations at the value of two (some young adolescents have smoked it a couple of times), and so on. Such skewed distributions are typical of many, but not all, count outcomes. The large positive skewness of counts can affect the accuracy of significance tests and confidence intervals if standard OLS regression methods are applied to the data. Statisticians have developed what they call count regression models to deal with such scenarios. More generally, these models are called **discrete regression models**.

Sometimes count data approximate normal distributions or are distributed in ways that traditional robust methods of regression can be applied to them with good effect. However, if the observed counts are near zero, then these approaches can yield predicted counts that are negative, a property that is bothersome to some methodologists (much like my discussion of negative probabilities for the MLPM). If your focus is on estimating coefficients in a causal framework (rather than generating individualized predicted outcomes in a prediction context), then the occurrence of negative counts may be less problematic (Angrist & Pischke, 2009). However, for highly skewed data with a large number of zeros, many methodologists prefer the use of specialized regression methods.

The Poisson Distribution and Count Data

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 share common features. For example, the mean of a Poisson distribution always equals its variance. The formal representation of Poisson distributions is provided in Long (1997) and I do not delve into its mathematical underpinnings here. Figure 5.5 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 when it has a mean of 10 or greater.

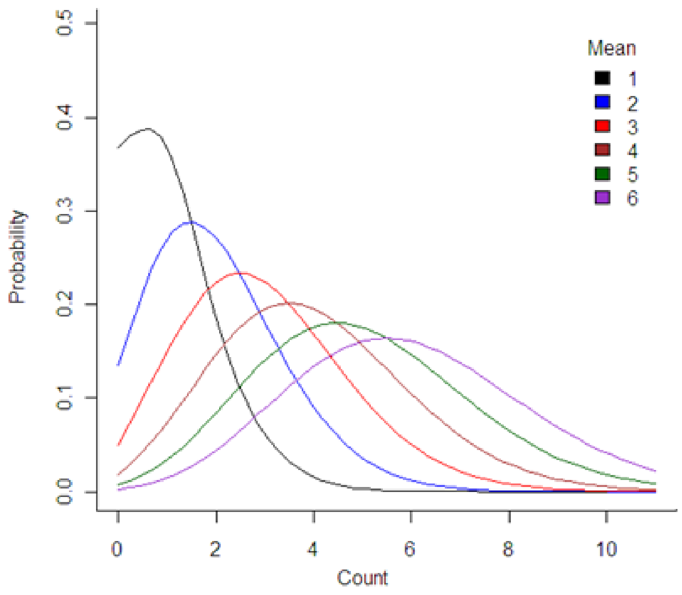

FIGURE 5.5. Poisson distributions

Look at the Poisson distribution in Figure 5.5 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 young adolescents. So, perhaps marijuana use in your data is Poisson distributed. Certainly, this is a more reasonable assumption than that the scores are normally distributed. If you are working with a count outcome that you think might be Poisson distributed, an informal way to determine this is to calculate its mean and variance. For a Poisson distribution, they should be equal. If the variance is larger than the mean, then the distribution is said to be **over-dispersed**. If the variance is smaller than the mean, the distribution is said to be **under-dispersed**. Of course, in sample data, the mean and variance may not be exactly equal due to sampling error. There are formal significance tests of over-dispersion and under-dispersion to take sampling error into account.

In addition to assuming Poisson distribution dynamics when we conduct a Poisson regression, a second assumption we make focuses on continuous predictors in the equation (or predictors that are discrete quantitative with many values and treated as if they are continuous). Consider the number of times adolescents between the ages of 12 and 17 have smoked marijuana in the past 30 days as a function of age. 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 these 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. Instead, Poisson regression assumes that the means are related to age by a log function, specifically:

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

As an example, suppose the mean number of times youth have smoked marijuana as a function of age are as follows (with natural logs of the means on the right):

Age	Mean Count (μ_i)	$\ln(\mu_i)$
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 because most adolescents, regardless of age, have not smoked marijuana; the data are dominated by zeros. Examine the column for the log of the μ. The entries for this column are 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. Although this relationship is linear, the relationship between age and the mean count per se is non-linear. For example, there is less change in the mean counts for a unit age increase at the lower end of the age distribution than at the upper end. A one-year increase in age from ages 12 to 13 is smaller than the corresponding increase from ages 16 to 17. Decisions to use Poisson regression also hinge on the function linking mean counts to continuous predictor values, a function that should be logarithmic in form.

Robust and Quasi-Likelihood Poisson Regression

It is possible that the Poisson model might describe the mean structure of Y across X profiles, but the distribution of the Y scores might not quite be Poisson distributed at the different X profiles. Perhaps the Y scores are close to being Poisson distributed, but there is some over-dispersion present. One strategy for dealing with this case is to proceed with the Poisson regression model but to use a robust estimator for the standard errors so that the significance tests and confidence intervals are not affected by over- or under-dispersion. This approach is called **robust Poisson regression** and the robust standard errors typically are based on Huber-White estimation. Yet another strategy for dealing with under- or over-dispersion is to use the Poisson model, but to modify it in a way that includes a parameter to adjust for it. This approach is called **quasi-likelihood Poisson regression**. I do not delve into quasi-likelihood Poisson regression here; see Cameron & Trivedi (1998) for details.

The Negative Binomial Distribution and Count Data

An alternative distribution to the Poisson distribution is the negative binomial distribution. When it is invoked, we conduct what is known as **negative binomial regression**. The negative binomial distribution also is a family of distributions that has different forms. There are two key parameters that impact its form, a mean and theta, the latter of which also is called the **shape parameter**. For a technical description of theta and the mathematics of the negative binomial distribution, see Long (1997). In a negative binomial distribution, a mean does not necessarily equal its variance, as was the case for a Poisson distribution. Figure 5.6 presents examples of negative binomial distributions with different means ranging from 1 to 6 and theta = 1; Figure 5.7 presents examples with a mean of 2 and thetas from 1 to 6. The negative binomial distribution is sometimes called the **Pascal distribution** or the **Pólya distribution**, which are special cases of it.

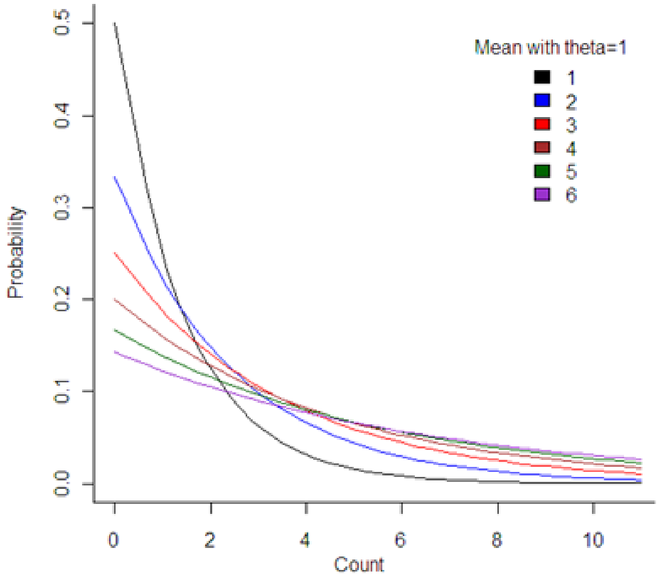

FIGURE 5.6. Negative binomial distributions with different means and theta = 1

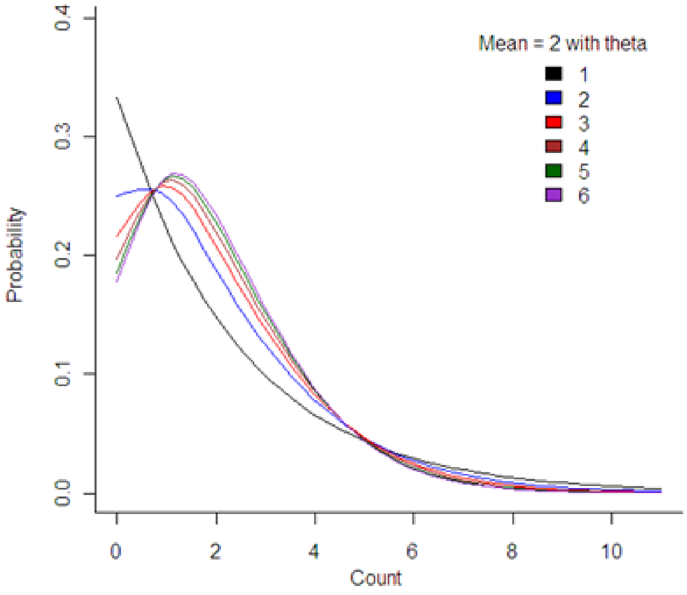

FIGURE 5.7. Negative binomial distributions with different thetas and a mean of 2

When we apply negative binomial regression, we assume the Y scores are distributed in accord with a negative binomial distribution at a given predictor profile. It turns out that across the values of a continuous predictor, X, negative binomial regression, like Poisson regression, assumes that ln(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, either Poisson or negative binomial.

Comparing Observed and Expected Frequencies

To help choose between the use of a Poisson and negative binomial regression model, we can compare predicted and observed count distributions based on the two regression approaches. For example, for 1,000 youth, I might find that 730 of them have never tried marijuana, 122 of them have tried marijuana once, 55 have tried marijuana twice, and so on. I refer to this as the **observed frequency distribution** for the count values. Using methods described by Long (1997), I can calculate a predicted frequency for each count based on the regression model I fit to the data. The mathematics of doing so are complex, but as you will see in future chapters, Mplus does the calculations for you. I call this distribution the **predicted frequency distribution** for the counts. If the Poisson regression model is appropriate, there should be close correspondence between the predicted and observed frequency distributions when I use Poisson regression. If the negative binomial model is appropriate, there should be close correspondence between the predicted and observed frequency distributions when I use negative binomial regression.

Figure 5.8 presents a plot of the observed and predicted (or expected) 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, the number of 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 5.9. These results support the choice of a negative binomial model as there is close correspondence between the predicted and observed count frequencies. Note that just because there is close correspondence between the predicted and observed frequency distributions does not mean the predictors are strongly related to the outcome at the level of individuals. The extent to which this is the case depends on the amount of variability in the counts at a given predictor profile. I discuss this issue in greater detail in future chapters.

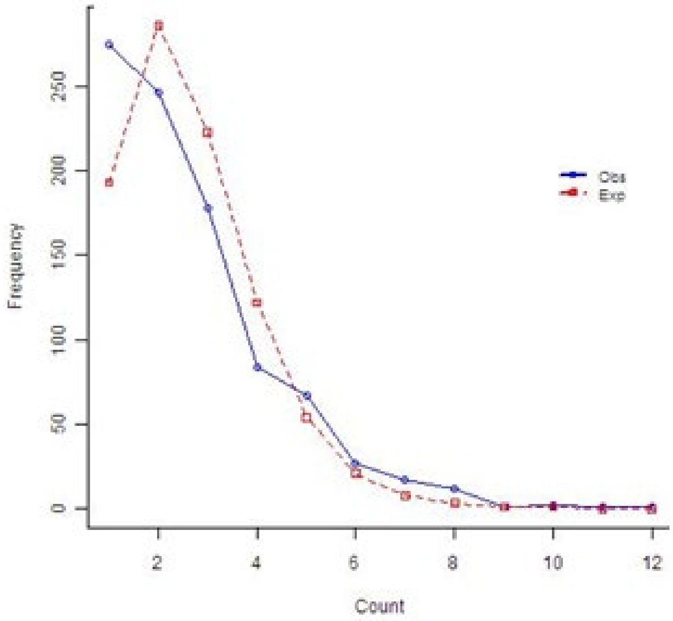

FIGURE 5.8. Poisson model observed and expected counts

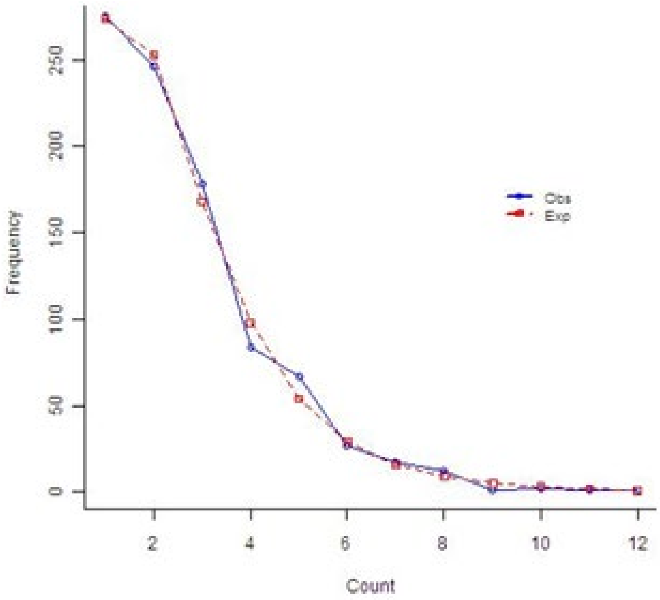

FIGURE 5.9. Negative binomial model observed and expected counts

Additional Plausible Count Distributions

My discussion thus far has focused on Poisson and negative binomial regression models. In this section, I briefly discuss other possible discrete regression models that can be used that also are available in Mplus.

Zero Inflated Regression Models

A possible source of over-dispersion in Poisson models is an excess of zeros produced by some theoretical mechanism other than those that derive from 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.

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

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

As an example, 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 undoubtedly will 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 following the above logic, and (b) there is close correspondence between the predicted and observed frequency distributions of the counts. 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 of zero inflated models, see Long (1997), Long and Freese (2006), and Cameron and Trivedi (1998, 2005).

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 mechanisms that underlie the inflation of zeros are different than traditional 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. For example, when predicting the number of times an adolescent has smoked marijuana, whether adolescents smoke marijuana for the first time might be impacted by somewhat different dynamics than how often they do so once they have started using marijuana. Unlike the aforementioned zero-inflation models, there are not two sources of zeros; the count dynamics are only activated if the zero hurdle is crossed and there is no "going back" once the hurdle has been crossed. As with the zero-inflation models, the same or different predictors can be used in each component. For details, see Cameron and Trivedi (1998, 2005).

Zero Truncated Regression 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 & Trivedi, 1998, 2005, for details.

Robust Linear Regression and Count Data

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 can be misspecified. In such situations, one might apply basic OLS regression to count data, but use robust algorithms, such as bootstrapping or a Huber-White sandwich estimator. As noted, if counts are near zero, then these approaches can yield predicted counts that are negative. 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 regression analysis is less problematic; see Angrist & Pischke, 2008, for elaboration of this point.

Thinking about data as constituting a count that requires specialized modeling is not always straightforward. For example, monthly income measured in dollars can be considered a count variable that reflects the number of dollars you earn over a 30 day period, but we rarely analyze income using count models or think of income as a count. Rather, we apply traditional regression modeling to income but perhaps with outlier corrections. You should not assume that just because a variable is a count, you must use discrete regression modeling methods. Many "count" variables can be analyzed appropriately using traditional regression methods.

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. We are 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 the count divided by the exposure. If we hold the exposure value constant at some value, E, then the average rate for predictor profile *i* is

$$\ln(\mu_i / E) = \alpha + \beta X_i \qquad [5.17]$$

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

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

Applying this to the left-hand side of Equation 5.17, yields

$$\ln(\mu_i) - \ln(E) = \alpha + \beta X_i$$

If I then add ln(E) to both sides of the equation, I get

$$\ln(\mu_i) = \alpha + \beta X_i + \ln(E)$$

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 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 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 (Cameron & Trivedi, 2005). 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 as well. 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 discrete regression models with offsets is most appropriate with Poisson regression. In my experience, Poisson models rarely are applicable in practice, so this limits the offset strategy described here. Rate data usually needs to be analyzed using other regression strategies than discrete regression modeling.

Interpreting Coefficients in Count Models

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

$$\ln(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

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. It is common practice to "remove the logs" by calculating exponents (or anti-logs) of the intercept and of the estimated β, much like we do in logistic regression.

Consider the estimated coefficients for the negative binomial model that predicted the number of publications of professors. Gender was treated as a dummy variable coded 1 for females and 0 for males. The exponent of its coefficient was 0.82. This value reflects the estimated mean count for the group scored 1 on the dummy variable divided by the estimated mean count for the reference group. It means that the estimated mean number of publications by women was 0.82 that of men; that the female mean is the estimated number for males multiplied by a factor of 0.82. For the "continuous" 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 increases 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 reflected by the exponent of the coefficient associated with X.

When you conduct a Poisson or negative binomial count regression and interpret the output, your focus generally will be 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 the value of 1.0, then the coefficient is judged to be statistically significant. I illustrate the use of count regression analyses in RETs in Chapter XX.

CONCLUDING COMMENTS

Regression methods are widely used in the analysis of RETs. Among the more widely used regression models are traditional linear regression, binary regression, ordinal regression, multinomial regression and count or discrete regression. All of these modeling strategies can be applied to RETs in the context of the unified system of structural equation models (SEM), with or without latent variables. I will illustrate such applications in future chapters.

APPENDIX: LATENT RESPONSE MODELS FOR BINARY REGRESSION

In this appendix, I develop the bases and implications of the latent response model for logistic and probit regression. The material is of interest because it makes evident some of the limitations of logistic and probit regression that often are underappreciated. I explore the implications for RET analyses in Chapter XX.

I will use the logistic regression model to develop the core ideas. As noted in the main text, the traditional logistic model is represented by the equation

$$\ln[\text{Odds}(Y)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad [A.1]$$

and the latent response version of it is

$$y_i^* = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k + \varepsilon_i \qquad [A.2]$$

Here, I elucidate the link between the β in Equation A.1 and the γ in Equation A.2 coupled with the threshold value, τ, namely that point on y* where non-performance of the outcome switches to performance. For reasons that will become apparent shortly, I prefer to express the disturbance term in Equation A.2 using a notation scheme by Mood (2010) and Karlson (2015) that introduces a scaling parameter, *s*, that allows the variance of ε to vary from an a priori assumed value as a function of this scaling parameter, as follows:

$$y_i^* = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k + (s)(\varepsilon_i) \qquad [A.3]$$

Note that when $s = 1.0$, Equation A.3 is no different from Equation A.2; as *s* becomes larger than 1.0, then the variance of the disturbances increases; as *s* becomes less than 1.0, the variance decreases. The need for *s* will be apparent shortly.

One challenge of the latent response formulation in Equation A.3 is that y* is not directly measured and consequently, it has no metric. Is it scored from 1 to 10, from 1 to 100, from -5 to +5, or what? In classic multiple regression, the regression coefficients and error variances take on values based on the metrics of Y and the metrics of the Xs. However, because y* has no metric, the propensity model is under-identified. By under-identified, I mean that the model has too many unknowns relative to the data at hand, so it cannot be estimated. This is because there are an infinite number of solutions for the values of γ, τ, and ε. It is like me giving you the equation $a + b = 36$ and asking you to tell me what the values of *a* and *b* are. There are an infinite number of solutions and we do not know which ones to use.

It turns out, that if we fix the mean and variance of (*s*)(ε) to take on certain values, then this reduces the degree of under-identification because it reduces the number of unknowns. For example, if we set the mean of (s)(ε) to be zero (a common assumption in regression models), $s = 1$, and the variance of ε to 3.290, then we do not need to estimate any of these values. This is the strategy that the latent response model uses to reduce under-identification; it fixes these parameters at the values just mentioned. It also assumes the ε follow a standard logistic distribution, which is a bell-shaped curve, much like a normal distribution. For probit models, s also is assumed to equal 1.0, the variance of ε is fixed at 1.0 instead of 3.290, and the ε are assumed to be normally distributed. Why choose the value 3.290 for the variance of ε in the logistic model? This is done because the variance of scores in a standard logistic distribution is $\pi^2/3$ (which equals 3.290) and because it has several statistical properties that assist statistical inference.

The act of fixing the error variance, ε, at 3.290 in logistic regression or 1.0 in probit regression is important. In traditional regression modeling, the variance of ε is impacted by the metric of the outcome and can take on any non-negative value based on that metric. If the outcome is measured on a 1 to 100 scale, then the variance of ε will be different than if the outcome is measured on a 1 to 10 scale. One goal of traditional regression analysis is to estimate the magnitude of the variance of ε. By contrast, in the latent response logit/probit regression framework, the variance of the ε is fixed and it never changes. This is, to say the least, unorthodox. Fixing the error variance at 3.290 actually is a form of "standardization," but it is a form of standardization that is different from what we are familiar with.

So, to summarize to this point, for Equation A.3, the mean, variance and distribution of ε per se never changes; for logit models, the mean is zero, the variance is 3.209, and it has a standard logit distribution; for probit models, the mean is zero, the variance is 1.0, and it has a normal distribution. The value of s can vary, but these other features of ε are set in stone. The parameter s is essentially an adjustment factor that modifies the disturbance variance to reflect its true variance. In theory, it equals the ratio of the true standard deviation of the errors divided by the assumed standard deviation of the errors (the latter of which is 3.290 in logit regression and 1.00 in probit regression). In practice, the value of s is not knowable. However, we need s to make the latent response formulation "work."

Even with these assumptions, there still remain sources of under-identification in the model. The source of this additional under-identification is the threshold value, τ, and the intercept, γ_0. It turns out these parameters also cannot be simultaneously estimated. One of them has to be fixed at an a priori value. In logistic and probit models, some statistical software fixes the threshold value at zero and the intercept is estimated. Other software fixes the intercept at 0 and estimates the threshold value.[10] Once this source of under-identification is rectified, the model becomes estimable and, coupled with the other statistical assumptions, it can be shown that the γ in Equation A.3 will equal the β in Equation A.2, but with the qualifications noted below.

I now elaborate ramifications of this framework, some of which are fundamental. The first ramification is that the metric and variance of y* is determined by two factors, (1) the variance $(s)\varepsilon_i$, which is assumed to be 3.290 in the logistic case if s=1, and (2) the variance of the predicted scores derived from the portion of Equation A.3 represented by $\gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_k X_k$. I refer to the latter as $\hat{y}^*$, which are the predicted y* scores based on the weighted predictors, ignoring the disturbance term. It can be shown that

$$\mathrm{var}(y^*) = \mathrm{var}\,(\hat{y}^*) + \mathrm{var}(s\varepsilon\,) = \mathrm{var}\,(\hat{y}^*) + 3.290$$

Because the error variance is fixed at 3.290, if we add a predictor with a non-zero γ to the prediction equation, the variance and metric of y* must change because var $(\hat{y}^*)$ will change. This property is not true in traditional multiple regression where the metric of Y remains the same no matter what variables are in the equation. In the latent response model, the metric and variance of y* are moving targets that can shift as one adds predictors to the equation. Note also that we cannot directly observe the value of s, and because we force ε to have a fixed variance (3.290), the β in Equation A.1 actually estimate β/s and not γ (see Karlson, 2015, for a proof).

There are non-trivial implications of this property that can create analytic challenges for logistic regression. For example, in traditional regression, if a consequential predictor of Y is relatively uncorrelated with the other predictors in the equation, then whether you leave that predictor out or include it in the prediction equation will not much affect the coefficients of the other predictors. This is not true of logistic regression because if you include such a variable in the equation, it will, by definition, alter the metric of y*. More specifically, when a viable predictor is added to the equation predicting y*, this has the effect of improving predictability by reducing the value of s (which reduces the overall error variance), but the reduced value of s also has the effect of changing the logistic coefficients, β, because these actually are β/s (Karlson, 2015). This means that characterizations of odds ratios and multiplicative factors are model dependent, i.e., they are tied to the specific predictors included in the analysis. In this sense, odds ratios and multiplicative factors lack generalizability because they are

[10] It turns out that if one fixes the intercept to zero, the threshold value will equal the intercept value for the case where the threshold is fixed at zero, but they will be opposite in sign.

influenced by the covariates in the equation (Norton, Dowd, & Maciejewski, 2018). I show the concrete ramifications of this property for the analysis of RETs in Chapter 12.

Given the assumptions that must be made to deal with under-identification as well as some of the contorted statistical properties of the latent response framework, some methodologists prefer not to think of logistic or probit regression in such terms. These methodologists also tend to shy away from odds ratios and multiplicative factors. They instead work with logistic-derived or probit-derived probabilities and average marginal effects. Mediation analysis with binary outcomes sometimes uses the latent response framework, as I illustrate in future chapters. The main take-away for now is that for a binary outcome, we often model a latent continuous propensity underlying it, y*, but the metric of this variable is, uncomfortably, impacted by the particular predictors in the equation, making substantive interpretation of the coefficients associated with the predictors somewhat awkward. All of the above also is relevant for ordinal regression because standard ordinal regression models use logistic or probit regression in their formulation.