1

A CONCEPTUAL INTRODUCTION TO BIVARIATE LOGISTIC REGRESSION

So you want to research something really interesting?

Let's say you want to research something interesting and important, like why students drop out of school before completing their degree, why people choose to use illicit drugs, what predicts whether an individual will die from a particular cause, whether a citizen will vote, or whether a consumer will purchase a particular type of product.

How would you do it? To be sure, researchers have been examining these types of outcomes for as long as curious people have been using scientific methods. But if they are not using logistic or probit regression (or similar procedure), odds[1] are they are not getting the most from their data.

Throughout the book, I will use simple, intuitive examples from a range of disciplines to demonstrate important aspects of logistic regression. In addition, example data sets will be available on the book's website so that readers can further enrich their logistic regression experience!

What is logistic regression, the oddly named and often underappreciated type of regression that many researchers in the social sciences have rarely, if ever, heard of? Decades ago, I took statistics courses from people

[1]Pun completely intended.

who I think were (or still are) some of the smartest and best teachers and scholars of statistics I have ever met. Despite having taken courses in regression models, ANOVA, multivariate statistics, hierarchical linear modeling, structural equation modeling, and psychometrics, I found that logistic regression was not covered in psychology and many social sciences disciplines back then. Indeed, many of the classic, beloved textbooks I used as a graduate student and as an assistant professor (such as the fabulous texts on regression by Pedhazur and Cohen and Cohen, and the excellent multivariate text by Tabachnick and Fidel) failed to cover the issue back then.[2] Today most texts covering regression at the graduate level give at least a cursory introduction to the topic, and the latest revisions of the classic texts I mention above now also introduce readers to the topic.

In fact, had I not by quirk of fate ended up working as a statistician and research associate in a medical school for several years, taking epidemiology courses and working with health science researchers, I would probably not have been exposed to logistic regression in any meaningful way. Logistic regression, I discovered, is widely used outside the particular niche of the social sciences I was trained in. Researchers in the health sciences (medicine, health care, nursing, epidemiology, etc.) have been using logistic (and probit) regression and other precursors for a very long time. Unfortunately, because it is a quirky creature, researchers often avoid, misuse, or misinterpret the results of these analyses, even in top, peer-reviewed journals where logistic regression is common (Davies, Crombie, & Tavakoli, 1998; Holcomb, Chaiworapongsa, Luke, & Burgdorf, 2001).

So why do we need a whole book dedicated to the exciting world of logistic regression when most texts cover the topic? It is a creature separate and unique unto itself, complex and maddening and amazingly valuable—when done right. Just as many books focus on analysis of regression (ANOVA), ordinary least squares (OLS) regression, factor analysis, multivariate statistics, structural equation modeling, hierarchical linear modeling, and the like, my years of experience using and teaching logistic regression to budding young social scientists leaves me believing this is a book that needs to be written. Logistic regression is different enough from OLS regression to warrant its own treatise. As you will see in coming chapters, while there are conceptual and procedural similarities between logistic and

[2]Of course, that was a long time ago. We calculated statistics by scratching on clay tablets with styli by candlelight and walked uphill, in the snow, both ways to get to class. Well, the second part at least is true. It was Buffalo back before climate change . . . everything was covered with snow year-round and everything was, indeed, uphill no matter what direction you were going. Or so it seemed with the wind. But I digress. The point is that it was just a really long time ago.

OLS regression, and to other procedures such as discriminant function analysis (DFA), the mathematics "under the hood" are different, the types of questions one can answer with logistic regression are a bit different, and there are interesting peculiarities in how one should interpret the results.

In other words, it is not the case that logistic regression is just multiple regression with a binary dependent variable. Well, yes, it is that, on the surface, and conceptually. But it is much more. The more I use it, and the more I teach it, and the more I try to dig into what exactly those numbers mean and how to interpret them, the more I have discovered that this stuff can be seriously confusing, and complex, interesting, and powerful. And really fun.

To be clear, it is in no way just multiple (OLS) regression with a binary outcome. My goal in this book is to explore the fun things researchers can do with logistic regression, to explicate and simplify the confounding complexities of understanding what logistic regression is, and to provide evidence-based guidance as to what I think are best practices in performing logistic regression.

WHAT IS ORDINARY LEAST SQUARES REGRESSION AND ♦ HOW IS LOGISTIC REGRESSION DIFFERENT?

We will get into the mathematics of how logistic regression works in subsequent chapters. Right now, there are a few conceptual similarities and differences that we can address to orient the reader who is not deeply familiar with the two types of analyses. First, let's remember that OLS regression—what we will often call linear regression or multiple regression—is a solid and very useful statistical technique that I have frequently used since the late 1980s. This contrasting is in no way attempting to set up logistic regression as superior to OLS regression (and certainly not vice versa). Just like I cannot say a hammer is a favored tool over a drill, I cannot give preference to one regression technique over another. They both serve different purposes, and they both belong in a hallowed place inside the researcher's toolbox.

The primary conceptual difference between OLS and logistic regression is that in logistic regression, the types of questions we ask involve a dichotomous (or categorical) dependent variable (DV), such as whether a student received a diploma or not. In OLS regression, the DVs are assumed to be continuous in nature.[3] Dichotomous DVs are an egregious

[3]Although, in practice, measurement in OLS regression is not always strictly continuous (or even interval).

violation of the assumptions of OLS regression and therefore not appropriate. Without logistic regression, a researcher with a binary or categorical outcome is left in a bit of a pickle. How is one to study the predictors of illness if in fact we cannot actually model how variables predict an illness?

Over the years, I have seen kludgy attempts such as using *t*-tests (or ANOVA) to explore where groups differ on multiple variables in an attempt to build theory or understanding. For example, one could look at differences between people who contract a disease and those who do not across variables such as age, race/ethnicity, education, body mass index (BMI), smoking and drinking habits, participation in various activities, and so on. Perhaps we would see a significant difference between the two groups in BMI and number of drinks per week on average. Does that mean we can assume that those variables might be causally related to having this illness? Definitely not, and further, it might also be the case that neither of these variables is really predictive of the illness at all. Being overweight and drinking a certain number of drinks might be related to living in a certain segment of society, which may in turn be related to health habits such as eating fresh fruits and vegetables (or not) and exercising, and stress levels, and commute times, and exposure to workplace toxins, which might in fact be related to the actual causes of the illness.

No disrespect to all those going before me who have done this exact type of analysis—historically, there were few other viable options (in addition, prior to large-scale statistical computing, logistic regression was probably too complex to be performed by the majority of researchers). But let's think for a minute about this process. There are many drawbacks to the approach I just mentioned. One issue is that researchers can have issues with power if they adjust for Type I error rates that multiple univariate analyses require (or worse, they might fail to do so). In addition, using this group-differences approach, researchers cannot take into account how variables of interest covary. This issue is similar to performing an array of simple correlations rather than a multiple regression. To be sure, you can glean some insight into the various relationships among variables, but at the end of the day, it is difficult to figure out which variables are the *strongest* or most important predictors of a phenomenon unless you model them in a multiple regression (or path analysis or structural equation modeling) type of environment.

Perhaps more troubling (to my mind) is the fact that this analytic strategy prevents the examination of interactions, which are often the most interesting findings we can come across. Let us imagine that we find sex

differences[4] between those who graduate and those who do not, and differences in household income between those who graduate and those who do not. That might be interesting, but what if in reality there is an interaction between the sex of the student and family income in predicting graduation or dropout rates? What if boys are much more likely than girls to drop out in more affluent families, and girls are more likely to drop out in more impoverished families? That finding might have important policy and practice implications, but we are unable to test for that sort of interaction using the method of analysis described above. Logistic regression (like OLS regression) models variables in such a way that we get the unique effect of the variables, controlling for all other variables in the equation. Thus, we get a more sophisticated and nuanced look at what variables are uniquely predictive (or related to) the outcome of interest.

I have also seen aggregation used as a strategy. Instead of looking at individual characteristics and individual outcomes, researchers might aggregate to a classroom or school level. So then researchers might think they have a continuous variable (0–100% graduation rate for a school) as a function of the percent of boys or girls in a school and the average family income. In my opinion, this does tremendous disservice to the data, losing information and leading to potentially misleading results. In fact, it changes the question substantially from "what variables contribute to student completion" to "what school environment variables contribute to school completion rates." Further, the predictor variables change from, say, sex of the student to percent of students who are male or female, and from race of student to percent of students who identify as a particular race, from family socioeconomic status (SES) to average SES within the school. These are fundamentally different variables, and, thus, analyses using these strategies answer a fundamentally different question. Furthermore, in my own explorations, I have seen aggregation lead to wildly overestimated effect sizes—double that of the appropriate analysis and more. Thus, aggregation changes the nature of the question, the nature of the variables, and can lead to inappropriate overestimation of effect sizes and variance accounted for.

---

[4]Readers may be more used to reading "gender differences" rather than "sex differences"—an example of American Psychological Association style and language use betraying the meaning of words—similar to the use of "negative reinforcement" as a synonym for punishment when in fact it is not at all. I will use the term "sex" in this book to refer to physical or biological sex—maleness or femaleness. Gender, conversely, refers to masculinity or femininity of behavior or psychology. The two concepts are not synonyms, and it does harm to the concepts to conflate them (Mead, 1935; Oakley, 1972). Please write your political leaders and urge them to take action to stop this injustice!

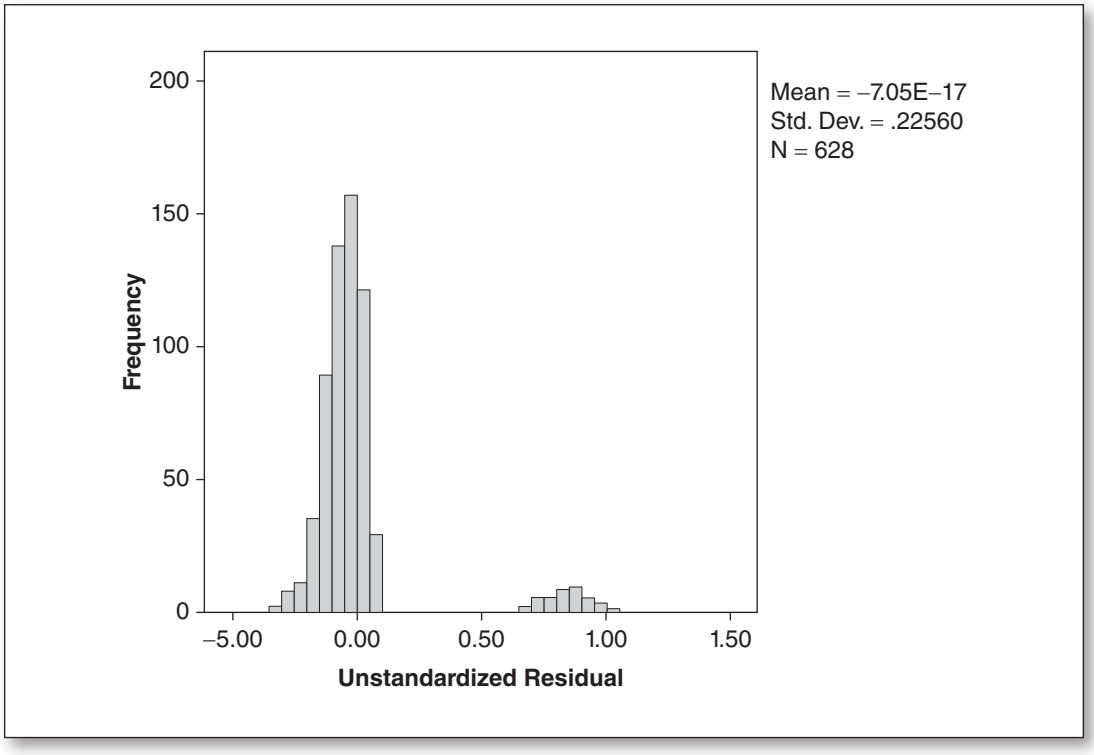
I am sure some of you have also wondered why we cannot just compute an OLS regression equation with a binary outcome as the DV. This is a real procedure often discussed in older regression texts and is referred to as the *linear probability model,* but it is not the same as a probit model (which I will cover later). This would carry the advantages of being able to simultaneously estimate the unique effects of several independent variables (IVs) and examine relative importance in predicting the outcome, unlike the approach described above. In fact, the statistical software you use will perform this analysis if you tell it to. But there are issues with this approach. First, predicted scores (which are supposed to be predicted probabilities) can range outside the acceptable 0.00–1.00 range. Second, residuals can only be 0, −1, or 1. Thus, they are neither normally distributed nor homoscedastic. In short, and without getting into too much detail, this simply is not an appropriate analysis (Cohen, Cohen, West, & Aiken, 2002). To illustrate this issue, I used a small subset of data from the National Education Longitudinal Study of 1988 (Ingels, 1994) to predict student completion (not dropping out) from some simple variables such as race, student grade point average, and student behavior problems. We will get back to this example data set later. For now, you can see in Figure 1.1 that performing this analysis in an OLS framework produced the expected violation of assumptions. For example, the residuals[5] are not close to being normally distributed.

And thus, we come to conceptual similarities between OLS and logistic regression. Procedurally, both OLS and logistic regression are set up with a single DV and one or more IVs. Both allow us to simultaneously assess the unique effects of multiple predictor IVs (and their interactions or curvilinear components, if desired), and both allow examination of residuals for purposes of screening data for outliers, follow-up analyses, or testing of assumptions. Both can perform simultaneous entry, hierarchical or blockwise entry (groups of IVs entered at one time), and various stepwise procedures.[6] And with both, we have the ability to assess a group of IVs to

[5]What is a residual? A residual is many things to many people—an error of estimation, error variance, unexplained variance, the unique person effect, the distance from the regression line to the data point, $Y - \hat{Y} \ldots$ All this is to say that however you interpret it, functionally it is the difference between the predicted value for an individual—their score predicted by the regression line equation—and their actual score.

[6]Many of you reading this will have been trained to have visceral negative reactions to stepwise procedures, which is ironic as just a generation earlier, they were heralded as important tools. I am of the mind that stepwise procedures have their place in the pantheon of statistical tools and that we should be knowledgeable of them and use them *when appropriate*. For most of you reading this, the answer to when these procedures are appropriate is "almost never." Indeed, a full discussion of stepwise procedures is beyond the scope of this book, but interested readers can refer to standard references for regression such as Cohen et al. (2002) and Pedhazur (1997).

Figure 1.1 Residuals From an OLS Analysis With Binary Outcome



Data Source: National Education Longitudinal Study of 1988 (NELS88), National Center for Educational Statistics (http://nces.ed.gov/surveys/nels88/).

determine which predictor is the strongest unique predictor of a particular outcome and to answer many of the types of questions that have made regression a valuable tool in quantitative methods.

OLS Regression—A Deeper Conceptual Look

Why do we call regression "ordinary least squares regression"? The ordinary least squares part refers to both the goal of the procedure and how it is calculated—the estimation method. The goal for OLS regression is to fit a straight line to the bivariate (or multivariate) scatterplot such that the line fits the data in the best way possible. Obvious, right? We would rarely want our regression line to be an inferior or misleading fit. So we have a "line of best fit" that we use as a single descriptor of the entirety of

the data we are analyzing. But how does that line get placed? Well, decades ago when statisticians and mathematicians were inventing this procedure, the obvious choice to them was to place the line such that the residuals were minimized. This goal of minimizing residuals—minimizing the distance between the data and the line of best fit—is intuitive and appealing in many ways. If assumptions are met and the line is fit well, most regression analyses will produce as many negative residuals as positive residuals (because as many data points will be below the line as above the line), and they generally sum to 0.00 or some value reasonably close. To get around this issue, statisticians use the simple step of squaring each residual and then summing them to get a positive value called the "sum of squares." This removes the issue of negative and positive residuals and gets closer to the idea of raw distance, as $-2^2 = 4$ and $2^2 = 4$. Research has shown that when assumptions are met in OLS regression that (a) the estimates produced are unbiased estimates of true regression properties within the population, (b) the standard errors decrease as sample size increases, and (c) they are efficient, meaning that no other method of estimation will produce smaller standard errors (if you are interested in more on the technicalities of OLS estimation, an excellent introduction is contained in Cohen et al., 2002).

Note particularly the phrase above "when assumptions are met." Too often in research we do not know if assumptions have been met because authors do not report having tested them. In fact, I wrote an entire book on why cleaning data and testing assumptions is so important (Osborne, 2012), and regression texts such as Cohen et al. (2002) clearly make the point that when assumptions are not met (e.g., the presence of even a single extreme outlier), bad things can happen to analyses (see particularly Cohen et al., 2002, Chapter 10, or Osborne, 2012—a most excellent book, in my opinion). More on assumptions in a few paragraphs.

Maximum Likelihood Estimation—A Gentle but Deeper Look

Maximum likelihood estimation is one of those developments in statistics that has spread primarily thanks to widespread access to statistical computing. Unlike OLS estimation, which is based on set equations that researchers or software can use to arrive at a calculated solution, maximum likelihood estimation is an *iterative* procedure. In other words, the software selects starting values for coefficients, calculates a solution, and compares it with a criterion. If the solution and the criterion are farther apart than

desired, new values are attempted and a new solution is found. Again, the new solution is examined and if found lacking, again is adjusted and a third solution is attempted. Hopefully, with each iteration, the solution approaches the goals of the algorithm. At some point, the last iteration will be accepted as the final estimation of effects, and that is what the researcher will see in the output. You can imagine how computationally intensive this procedure is and why it was not widely used until computing power became widely available.

Without getting into too many technical details, the goal of maximum likelihood estimation (MLE) is to find a solution that provides intercepts and slopes for predictor variables that *maximizes* the *likelihood* of individuals having scores on the dependent variable (Y) given their scores on the predictor variables (X_1, X_2, etc.). In other words, the algorithms are maximizing the likelihood that we would obtain the sample—the data, the observed scores—given the model and parameters being estimated. We have observed scores on variables for individuals within the sample that arose from some real dynamic or relationship within the population. The MLE algorithm attempts to provide a model that maximizes the likelihood of producing the results observed. In essence, both OLS and MLE are attempting to summarize the observed data. The two procedures are merely using different mathematical techniques to get to that goal.

Maximum likelihood estimation is, in my mind, similar to the somewhat counterintuitive notion of hypothesis testing and p values. The actual interpretation of a p value is the probability of obtaining the observed data if in fact the null hypothesis (H_0) is true in the population.[7] So conceptually, what MLE is trying to do is to estimate the various parameters (slopes and intercepts) that best model (or re-create) the observed data. Thus, if we have a population wherein the height of women and their shoe sizes are strongly positively related (as evidenced by the observed data), MLE will provide the coefficients and slopes that maximize the likelihood of obtaining the observed sample that contains the observed relationship between height and shoe size. MLE will repeatedly attempt estimations based on slightly different coefficients until the fit with the observed data is as good as can be—in other words, that successive iterations fail to improve the fit by an appreciable amount.

[7]It is not, contrary to popular belief, the probability of being wrong, the probability of getting the observed results by random chance, etc. It is also not exactly what we really want to test—which is the probability that our alternative hypothesis (H_a) is true.

♦ DIFFERENCES AND SIMILARITIES IN ASSUMPTIONS BETWEEN OLS AND LOGISTIC REGRESSION

Distributional Assumptions

Because MLE has different mathematical estimation than OLS, MLE has some different assumptions than OLS. OLS regression is a parametric technique, meaning that it requires assumptions about the distribution of the data in order to be effective (these are discussed in most regression texts, but a particularly good reference is Cohen et al., 2002; see also Osborne & Waters, 2002). Commonly used statistical tests such as ANOVA and OLS regression assume that the data come from populations that are normally distributed or that have normal distributions of residuals (errors). In contrast, because of the different estimation procedures, logistic regression is a *nonparametric* technique, meaning it does not require any particular distributional assumptions.

Linearity of the Relationship

Another assumption of OLS regression is often referred to as the "assumption of linearity." The general assumption is that the correct form of the relationship is being modeled, but in the case of OLS regression and many other analyses, the assumption is that there is a linear relationship between the DV and the IV. A similar generalization to planes and hyperdimensional relationships is in effect for multiple regression with 2 or more IVs, but thinking too deeply about hyperdimensional generalizations of linearity gives me a bit of a headache, so I tend to stick to the 2- or 3-dimensional examples. Interestingly, I have often found relationships that are curvilinear in nature, not only in the social sciences but also in health sciences. Immediate examples that come to mind can include the relationship between arousal (i.e., stress) and performance (Loftus, Loftus, & Ketcham, 1992; Sullivan & Bhagat, 1992; Yegiyan & Lang, 2010),[8] student achievement growth curves (Francis, Schatschneider, & Carlson, 2000; Rescorla & Rosenthal, 2004), grade point average and employment in high school students (Quirk, Keith, & Quirk, 2001), dose-response relationships

[8]This is often attributed to Yerkes-Dodson (Yerkes & Dodson, 1908), or somewhat inaccurately referred to (sometimes by me personally) as an anxiety-performance curve. See Teigen (1994) for a historical overview of this large group of theories and studies.

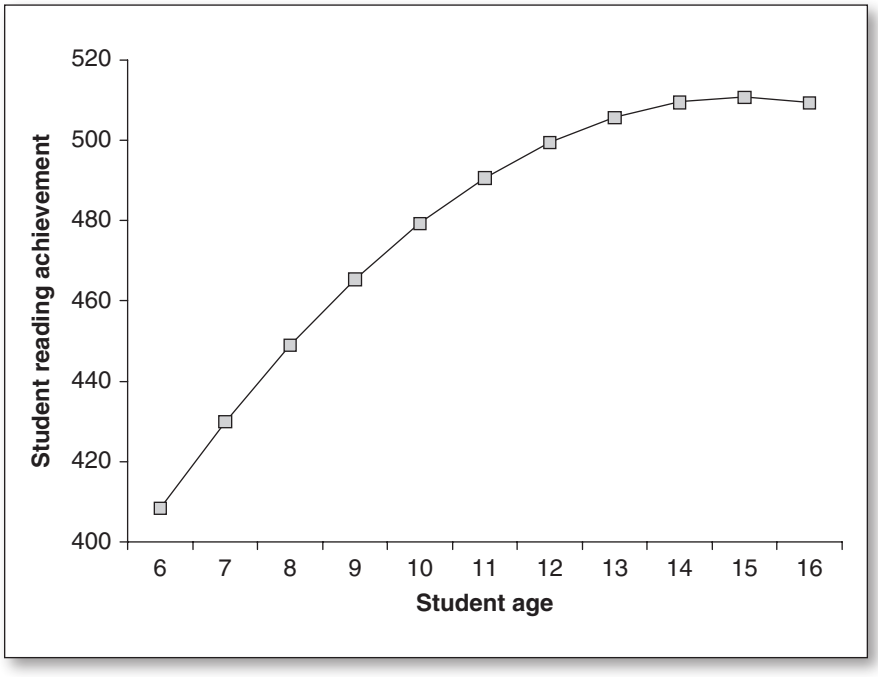
(Davis & Svendsgaard, 1990), and age and life satisfaction (Mroczek & Spiro III, 2005).

For example, as Francis et al. (2000) showed, the general pattern for reaching achievement growth over time is curvilinear. In Figure 1.2, I present a growth curve modeled from their published data.

When this assumption of linearity is violated, two things happen. First, really interesting findings are overlooked, and second, OLS regression will underestimate (and mischaracterize) the true nature of the relationship. Fortunately, there are increasingly easy ways to incorporate tests for curvilinear effects as statistical software packages begin to implement curvilinear regression options.

Logistic regression is, by nature, nonlinear, as we will discuss in more detail in subsequent chapters. Specifically, the way that logistic regression converts a dichotomous or categorical variable to a dependent variable that

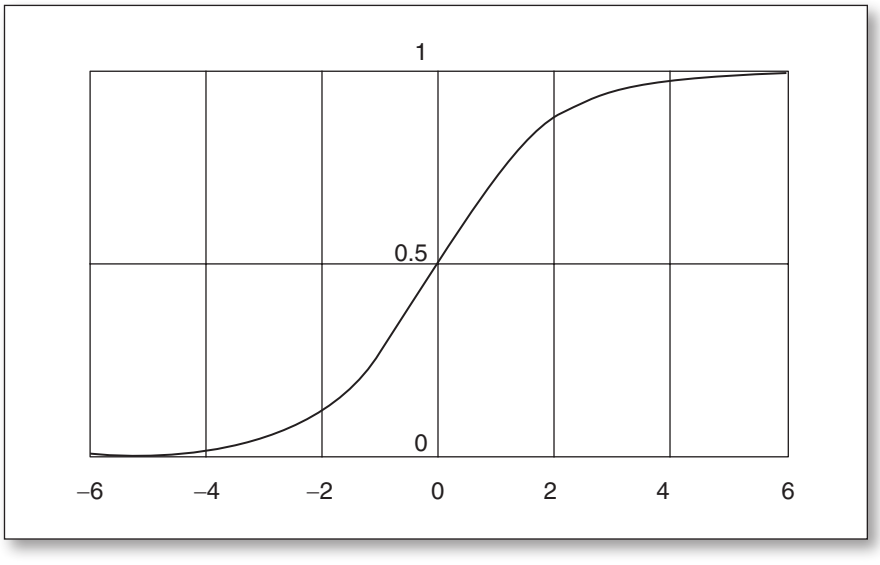
Figure 1.2 Curvilinear Relationship Between Student Age and Reading Achievement Test Scores



Data Source: Francis, D., Schatschneider, C., & Carlson, C. (2000). Introduction to individual growth curve analysis. In D. Drotar (Ed.), *Handbook of research in pediatric and clinical child psychology* (pp. 51–73). New York, NY: Kluwer/Plenum.

can be predicted from other binary, categorical, or continuous variables involves a nonlinear transformation. For now, envision a dependent variable that is an S-shaped curve representing the probability that an individual will be in one group or the other (like the one in Figure 1.3). Don't worry about the details of how the DV is created in logistic regression for now—we will have fun exploring that more thoroughly later. I find it interesting that although the basic character of logistic regression—the logit transformation—is curvilinear, there is a clear assumption of linearity as well. Specifically, there is an assumption that there is a linear relationship between IVs and the DV—that IVs are "linear on the logit."[9] Similar to when we create models in OLS regression, we can model relationships that are

Figure 1.3 Standard Logistic Sigmoid Function



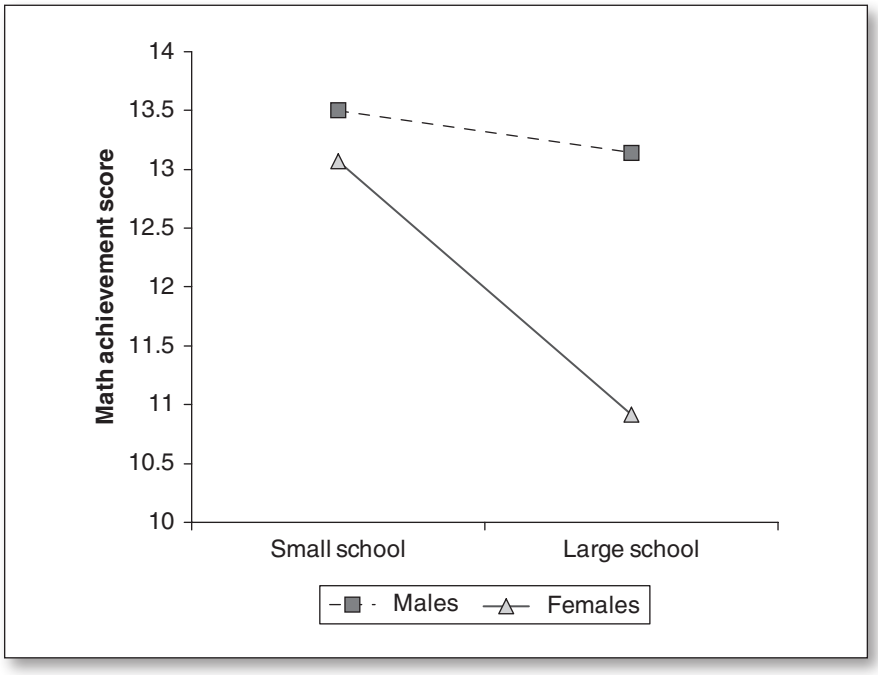
Source: Wikipedia (http://upload.wikimedia.org/wikipedia/commons/thumb/8/88/Logistic-curve.svg/320px-Logistic-curve.svg.png).

"nonlinear on the logit" easily, something we will explore in more depth in Chapter 7.

Note that "nonlinearity" includes the concept of interactions, where the effects of one variable depend on the effect of a second variable. For

[9]This is also a great phrase to drop casually into conversations. Try it and watch your social capital climb!

Figure 1.4 Relationship of School Size and Student Sex Predicting Math Achievement Test Score



Data source: High School and Beyond data (http://nces.ed.gov/pubsearch/getpubcats.asp?sid=022).

example, there has been a lot of discussion about sex differences in math achievement test scores. This was a particular issue back in the 1980s when the National Center for Educational Statistics began their High School and Beyond study of high school students (information and data from HS&B available at http://nces.ed.gov/pubsearch/getpubcats.asp?sid=022). In the 1980s, we knew that in general, girls underperformed on mathematics achievement tests compared with boys. As Figure 1.4 shows, that general pattern, however, is not the same in all types of schools. In this example,[10] we see that there is a trend for much larger schools to have larger math achievement gaps and for smaller schools to have smaller achievement gaps.

[10]Note this is not a logistic regression example—we will explore interactions in logistic regression in Chapter 8. These results are from a quick hierarchical linear modeling (HLM) analysis with a continuous DV (HLM will be introduced and discussed as it relates to logistic regression in Chapter 13).

In other words, the effect of student sex differs depending on the size of the school (and probably many other variables). Again, there are many reasons for this that we will not get into, but the conclusion is interesting—that context matters. It is not simply a linear relationship, but rather a variety of relationships dependent upon the context the students find themselves in. Similarly, it is likely that when you look for them, you will find and model interaction effects when running logistic regressions. We will explore interaction effects in more detail in Chapter 8.

Perfect Measurement

It is almost a dirty little secret in statistical science that we assume perfect measurement yet rarely achieve it. In most statistical procedures, we assume we are measuring the variables of interest well, and to the extent that we are not, biases and misestimations can occur. In simple correlation and regression, the effect is usually that of underestimating the effects in question. Yet in multiple regression and more complex procedures, the effects can get unpredictable and chaotic. For example, if you are studying student achievement and attempting to control student socioeconomic status (SES), but your measure of SES is imperfect, then you are failing to fully control for the effect. This can lead other, related IVs to become overestimated if they are capturing variance that should have been removed by SES. I have dealt with this issue in more depth in other places and so will refer interested readers to those rather than recapitulating the arguments here (Nimon, Zientek, & Henson, 2012; Osborne, 2003, 2008, 2012). Logistic regression also relies upon reliable measurement of variables, and so in that respect the two are similar.

Homoscedasticity (or Constant Variance of the Residuals)

ANOVA has the assumption of equal variance across groups, and OLS regression has a similar assumption—that the variance of the residuals (or the variance of data points around the regression line) is constant across the observed range. In other words, this means that if you plot the data points around the regression line (e.g., a ZPRED vs. ZRESID plot), you should see a relatively homogenous scattering of data points around the regression line at all points.

Again, because logistic regression is not a parametric procedure, there is no assumption of homoscedasticity or equality of variance. But there are some interesting assumptions relating to sparseness that seem similar to me.

Sparseness is a concept that can be understood by imagining lots of little boxes stacked together. Each box represents a combination of the DV and IV. For example, if you are looking at blood pressure and odds of having a stroke, you have boxes for each range of blood pressure representing both people who have and who have not had strokes. In sampling from the population, you want to make sure you have your boxes filled as best you can. "Sparse" data refers to having some of these boxes unfilled or not filled enough to allow the MLE estimation to effectively form estimates. This is an interesting difference between OLS and logistic regression and we will examine assumption testing in more detail in Chapter 4.

Independence of Observations

In most analyses, we assume that observations are independent unless we are specifically modeling nested data or repeated measures. Because much of our data in the world (especially in the social sciences, but also in many other sciences, such as health sciences) comes from organisms that form hierarchies or groups, this assumption may be more or less tenable. For example, researchers sampling individuals from existing health centers or students from schools or classrooms are sampling individuals who are already more similar in many respects than individuals sampled at random from the entire population. This violates the assumption of independence of observation and may bias the results. For a brief primer on this concept, and the issues that can arise, you can refer to Osborne (2000) or refer to Chapter 13, where we discuss HLM applied to logistic regression.

SIMILARITIES BETWEEN OLS ♦ REGRESSION AND LOGISTIC REGRESSION

Summarizing the Overall Model

One of the first things many researchers look at in OLS regression is the overall model fit, usually represented with a multiple *R,* with associated significance test, and R^2, the overall amount of variance accounted for. This is an important statistic and represents goodness of the model. Logistic regression does not have an exact analogue to R^2. Instead we have the concept of *deviance,*[11] which represents lack of fit or deviance from

[11]Which is much less exciting in discussing logistic regression than in discussing social or behavioral deviance. Sorry.

the observed data. In logistic regression, we can start with deviance for the null model, or the overall amount of deviance—essentially the overall amount of deviance that can potentially exist in the dependent variable. Then we have model deviance, the deviance that remains once predictor variables have been added to the model. Deviance is reduced as significant predictors are added, and there are statistical tests for this reduction similar to that of ΔR^2, usually in the form of a χ^2 that is the difference between the null or baseline model and the final model, with degrees of freedom that represent the number of parameters estimated that changed between the two models. This test is called the *likelihood ratio test* (Hosmer & Lemeshow, 2000).

$$Deviance = -2(\ln \frac{\text{likelihood of fitted model}}{\text{likelihood of saturated model}}) \qquad \text{Eq. 1.1.}$$

So conceptually, there are ways to assess the overall model in logistic regression, but the method differs significantly in terms of what deviances are and how they are thought of. If you are familiar with other types of analyses that use maximum likelihood, you may have seen −2 log likelihood used similarly. Deviances and −2 log likelihoods are conceptually identical.

♦ WHAT IS DISCRIMINANT FUNCTION ANALYSIS AND HOW IS LOGISTIC REGRESSION SUPERIOR/DIFFERENT?

I briefly discussed the idea of performing an OLS regression analysis with the binary DV. This analysis is referred to as the linear probability model, and to recap, there are multiple issues with this approach. For example, predicted probabilities can exceed the 0.00 to 1.00 range that is conceptually valid; the residuals are highly heteroscedastic and not normally distributed. Two-group discriminant analysis was developed early in the 20th century (Fisher, 1936). In practice, this procedure was often used to classify individuals based on certain predictor variables to explore whether a researcher could account for, say, a clinician's diagnosis. In discriminant analysis, a set of predictors is used to generate a prediction equation, called the linear discriminant function, with each predictor weighted with a coefficient (just as in OLS regression), and predicted scores. While somewhat intuitive, discriminant analysis is mathematically identical to the linear probability model (Cohen et al., 2002) and thus carries the same liabilities. Thus, it is considered an anachronistic

procedure and does not currently represent a best practice. Instead, researchers should use logistic regression, which is considered the successor to this procedure.

SUMMARY

Logistic regression is a relative newcomer to the statistical toolbox, particularly in the social sciences, but it is currently considered *the* best practice when dealing with outcomes that are dichotomous or categorical in nature. Through the course of this book we will explore all the various ways logistic regression is similar to, and different from, OLS regression. If you are familiar with OLS regression, you will find logistic regression a simple-to-understand cousin. The technical details "under the hood" are very different, and there are some interesting and fun nuances that an expert logistic regression user needs to master (but in fairness, there are many interesting and fun nuances that expert OLS regression users need to master as well). We will take each topic one at a time, and by the end it is my hope that you will appreciate the beauty and power of this procedure, ready to use it according to evidence-based best practices.

REFERENCES

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal, 316,* 989–991.

Davis, J. M., & Svendsgaard, D. J. (1990). U-Shaped dose-response curves: Their occurrence and implications for risk assessment. *Journal of Toxicology and Environmental Health, Part A Current Issues, 30*(2), 71–83.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188. doi: citeulike-article-id:764226

Francis, D., Schatschneider, C., & Carlson, C. (2000). Introduction to individual growth curve analysis. In D. Drotar (Ed.), *Handbook of research in pediatric and clinical child psychology* (pp. 51–73). New York, NY: Klewer/Plenum.

Holcomb, W. L., Jr., Chaiworapongsa, T., Luke, D. A., & Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics and Gynecology, 84*(4), 685–688.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Hoboken, NJ: Wiley.

Ingels, S. (1994). *National Education Longitudinal Study of 1988: Second follow-up: Student component data file user's manual*. Washington, DC: U.S. Department

of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Loftus, E. F., Loftus, E., & Ketcham, K. (1992). Witness for the defense: The accused, the eyewitness, and the expert who puts memory on trial. New York, NY: St. Martin's Griffin.

Mead, M. (1935). Sex and temperament in three primitive societies. New York, NY: Morrow.

Mroczek, D. K., & Spiro III, A. (2005). Change in life satisfaction during adulthood: findings from the veterans affairs normative aging study. *Journal of Personality and Social Psychology, 88*(1), 189.

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*(102).

Oakley, A. (1972). *Sex, Gender, and society.* London: Temple Smith.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation, 7*(1).

Osborne, J. W. (2003). Effect Sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation, 8*(99).

Osborne, J. W. (2008). Is disattenuation of effects a best practice? In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 239–245). Thousand Oaks, CA: Sage.

Osborne, J. W. (2012). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Thousand Oaks, CA: Sage.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation, 8*(2).

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction. Fort Worth, TX: Harcourt Brace College.

Quirk, K. J., Keith, T. Z., & Quirk, J. T. (2001). Employment during high school and student achievement: Longitudinal analysis of national data. *The Journal of Educational Research, 95*(1), 4–10.

Rescorla, L., & Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology, 96*(1), 85.

Sullivan, S. E., & Bhagat, R. S. (1992). Organizational stress, job satisfaction and job performance: Where do we go from here? *Journal of Management, 18*(2), 353–374.

Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. *Theory & Psychology, 4*(4), 525–547.

Yegiyan, N. S., & Lang, A. (2010). Processing central and peripheral detail: How content arousal and emotional tone influence encoding. *Media Psychology, 13*(1), 77–99.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459–482.