Chapter 1

THE LOGIC OF LOGISTIC REGRESSION

Many social phenomena are discrete or qualitative rather than continuous or quantitative in nature—an event occurs or it does not occur, a person makes one choice but not the other, an individual or group passes from one state to another. A person can have a child, die, move (either within or across national borders), marry, divorce, enter or exit the labor force, receive welfare benefits, vote for one candidate, commit a crime, be arrested, quit school, enter college, join an organization, get sick, belong to a religion, or act in myriad ways that either involve a characteristic, event, or choice. Sometimes continuous scales are measured qualitatively, such as for income below the poverty level or birth weight below a specified level. Likewise, large social units—groups, organizations, and nations—can emerge, break up, go bankrupt, face rebellion, join larger groups, or pass from one type of discrete state into another.

Binary discrete phenomena usually take the form of a dichotomous indicator or dummy variable. Although it is possible to represent the two values with any numbers, employing variables with values of 1 and 0 has advantages. The mean of a dummy variable equals the proportion of cases with a value of 1 and can be interpreted as a probability.

Regression With a Binary Dependent Variable

A binary dependent variable with values of 0 and 1 seems suitable on the surface for use with multiple regression. Regression coefficients have a useful interpretation with a dummy dependent variable—they show the increase or decrease in the predicted probability of having a characteristic or experiencing an event due to a one-unit change in the independent variables. Equivalently, they show the change in the predicted proportion of respondents with a value of 1 due to a one-unit change in the independent variables. Given familiarity with proportions and probabilities, researchers should feel comfortable with such interpretations.

The dependent variable itself only takes values of 0 and 1, but the predicted values for regression take the form of mean proportions or probabilities conditional on the values of the independent variables. The higher the predicted value or conditional mean, the more likely that any individual with particular scores on the independent variables will have a characteristic or experience the event. Linear regression assumes that the conditional proportions or probabilities define a straight line for values of the independent variables.

To give a simple example, the 2017 National Health Interview Survey asked respondents if they currently smoke cigarettes or not. Assigning those who smoke a score of 1 and those who do not a score of 0 creates a dummy dependent variable. Taking smoking (S) as a function of years of completed education (E) and a dummy variable for gender (G) with men coded 1 produces the regression equation:

$$S = .388 - .018E + .039G$$

The coefficient for education indicates that for a 1-year increase in education, the predicted probability of smoking goes down by .018, the proportion smoking goes down by .018, or the percent smoking goes down by 1.8. Women respondents with no education have a predicted probability of smoking of .388 (the intercept). A woman with 10 years of education has a predicted probability of smoking of .388 − (.018 × 10) = .208. One could also say that the model predicts 20.8% of such respondents smoke. The dummy variable coefficient for gender shows men have a probability of smoking .039 higher than for women. With no education, men have a predicted probability of smoking of .388 + .039 = .427.

Despite the uncomplicated interpretation of the coefficients for regression with a dummy dependent variable, the regression estimates face two sorts of problems. One type of problem is conceptual in nature, while the other type is statistical in nature. The problems may prove serious enough to use an alternative to ordinary regression with binary dependent variables.

Problems of Functional Form

The conceptual problem with linear regression with a binary dependent variable stems from the fact that probabilities have maximum and minimum values of 1 and 0. By definition, probabilities and proportions cannot exceed 1 or fall below 0. Yet, the linear regression line will continue to extend upward as the values of the independent variables increase, and continue to extend downward as the values of the independent variables decrease. Depending on the slope of the line and the observed X values, a model can give predicted values of the dependent variable above 1 and below 0. Such values make no sense and have little predictive use.

A few charts can illustrate the problem. The normal scatterplot of two continuous variables shows a cloud of points as in Figure 1.1(a). Here, a line through the middle of the cloud of points would minimize the sum of squared deviations. Further, at least theoretically, as X extends on to higher or lower levels, so does *Y*. The same straight line can predict large *Y* values associated with large *X* values as it can for medium or small values. The scatterplot of a relationship of a continuous independent variable to a dummy dependent variable in Figure 1.1(b), however, does not portray a cloud of points. It instead shows two parallel sets of points. Fitting a straight line seems less appropriate here. Any line (except one with a slope of 0) will eventually exceed 1 and fall below 0.

Some parts of the two parallel sets of points may contain more cases than others, and certain graphing techniques can reveal the density of cases along the two lines. For example, jittering reduces overlap of the scatterplot points by adding random variation to each case. In Figure 1.2, the jittered

Figure 1.1 (a) Scatterplot, continuous variables and (b) scatterplot, dummy dependent variable.

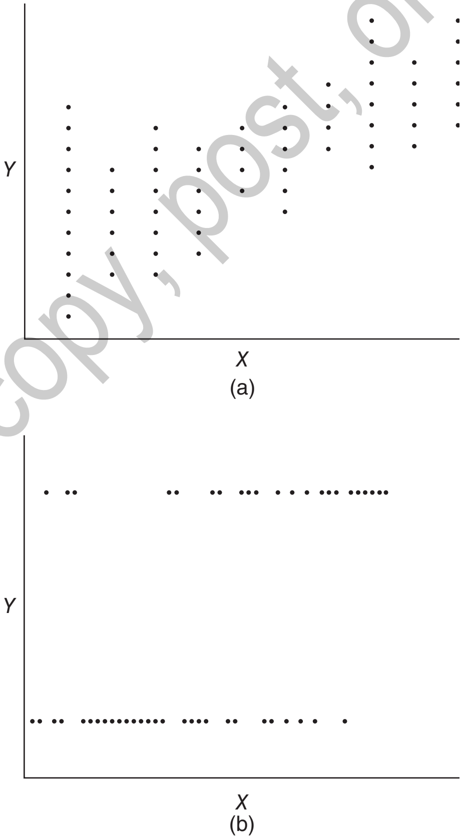
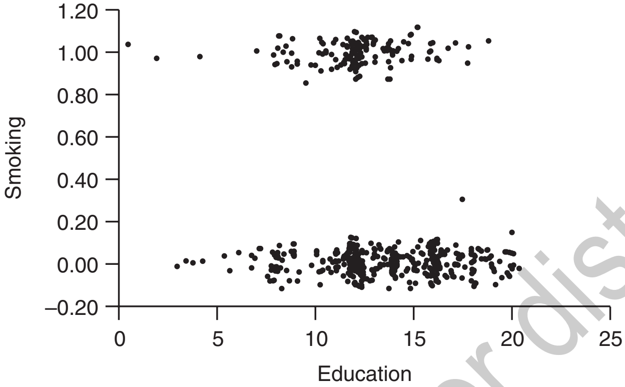


Figure 1.2 Jittered scatterplot for a binary dependent variable, smoking or nonsmoking, by years of education.



distribution for a binary dependent variable—smokes or does not smoke—by years of education suggests a slight relationship. Cases with higher education appear less likely to smoke than cases with lower education. Still, Figure 1.2 differs from plots between continuous variables.

Predicted probabilities below 0 or above 1 can occur, depending on the skew of the outcome, the range of values of the independent variable, and the strength of the relationship. With a skewed binary dependent variable, that is with an uneven split in the two categories, predicted values tend to fall toward the extremes. In the example of smoking, where the split equals 15:85, the lowest predicted value of .062 occurs for women with the maximum education of 18; the highest predicted value of .427 occurs for men with the minimum education of 0. However, simply adding age to the model produces predicted values below 0 for females aged 75 years and over with 18 years of education.

The same problem can occur with a less skewed dependent variable. From 1973 to 2016, the General Social Survey (GSS) asked respondents if they agree that the use of marijuana should be made legal. With the 30% agreeing coded 1 and the 70% disagreeing coded 0, a regression of agreement (*M*) on years of education (*E*), a dummy variable for gender (*G*) with males coded 1, and a measure of survey year (*Y*) with the first year, 1973, coded 0 and each year thereafter coded as the years since 1973, gives

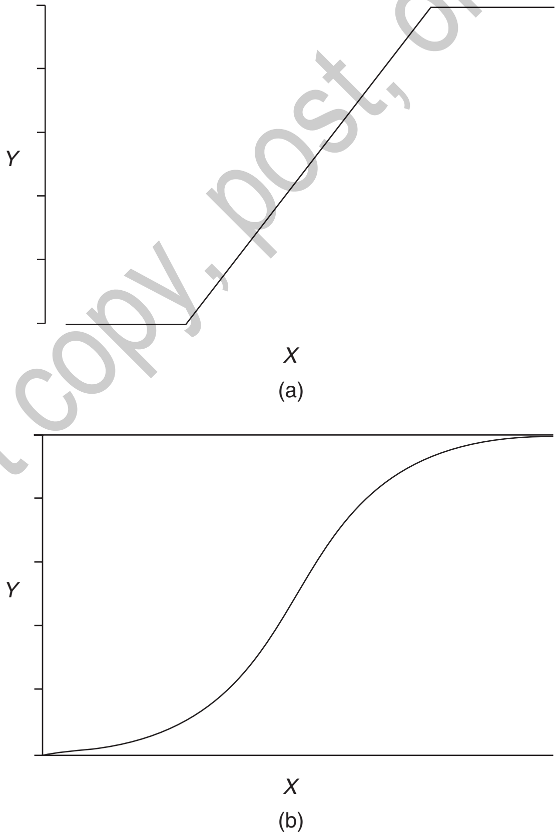
$$M = -.104 + .017E + .083G + .007Y$$

The intercept for females with no years of education and responding in 1973 shows the nonsensical predicted probability well below 0. Although a problem in general, reliance on the assumption of linearity in this particular model proves particularly inappropriate.[1]

Alternative to Linearity

One solution to the boundary problem would assume that any value equal to or above 1 should be truncated to the maximum value of 1. The regression line would be straight until this maximum value, but afterward changes in the independent variables would have no influence on the dependent variable. The same would hold for small values, which could be truncated at 0. Such a pattern would define sudden discontinuities in the relationship, whereby at certain points the effect of *X* on *Y* would change immediately to 0 (see Figure l.3(a)).

Figure 1.3 (a) Truncated linear relationship and (b) S-shaped curve.



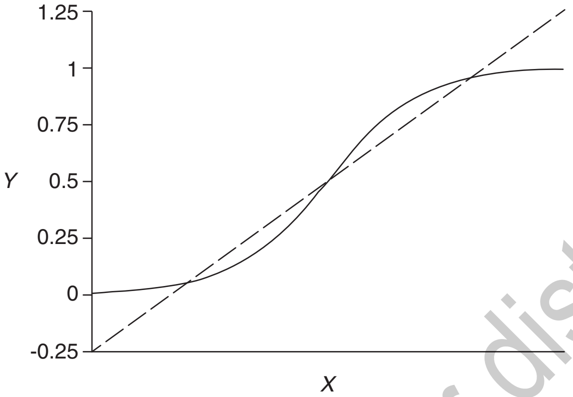
However, another functional form of the relationship might make more theoretical sense than truncated linearity. With a floor and a ceiling, it seems likely that the effect of a unit change in the independent variable on the predicted probability would be smaller near the floor or ceiling than near the middle. Toward the middle of a relationship, the nonlinear curve may approximate linearity, but rather than continuing upward or downward at the same rate, the nonlinear curve would bend slowly and smoothly so as to approach 0 and 1. As values get closer and closer to 0 or 1, the relationship requires a larger and larger change in the independent variable to have the same impact as a smaller change in the independent variable at the middle of the curve. To produce a change in the probability of the outcome from .95 to .96 requires a larger change in the independent variable than it does to produce a change in the probability from .45 to .46. The general principle is that the same additional input has less impact on the outcome near the ceiling or floor, and that increasingly larger inputs are needed to have the same impact on the outcome near the ceiling or floor.

Several examples illustrate the nonlinear relationship. If income increases the likelihood of owning a home, an increase of 10 thousand dollars of income from $70,000 to $80,000 would increase that likelihood more than an increase from $500,000 to $510,000. High-income persons would no doubt already have a high probability of home ownership, and a $10,000 increase would do little to increase their already high probability. The same would hold for an increase in income from $0 to $10,000: since neither income is likely to be sufficient to purchase a house, the increase in income would have little impact on ownership. In the middle-range, however, the additional $10,000 may make the difference between being able to afford a house and not being able to afford a house.

Similarly, an increase of 1 year in age on the likelihood of first marriage may have much stronger effects during the late twenties than at younger or older ages. Few will marry under age 18 despite growing a year older, and few unmarried by 50 will likely marry by age 51. However, the change from age 29 to 30 may result in a substantial increase in the likelihood of marriage. The same kind of reasoning would apply in numerous other instances: the effect of the number of delinquent peers on the likelihood of committing a serious crime, the effect of the hours worked by women on the likelihood of having a child, the effect of the degree of party identification on the support for a political candidate, and the effect of drinking behavior on premature death are all likely stronger at the midrange of the independent variables than the extremes.

A more appropriate nonlinear relationship would look like that in Figure l.3(b), where the curve levels off and approaches the ceiling of 1 and

Figure 1.4 Linear versus curvilinear relationship.



the floor of 0. Approximating the curve would require a succession of straight lines, each with different slopes. The lines nearer the ceiling and floor would have smaller slopes than those in the middle. However, a constantly changing curve more smoothly and adequately represents the relationship. Conceptually, the S-shaped curve makes better sense than the straight line.

Within the range of a sample, the linear regression line may approximate a curvilinear relationship by taking the average of the diverse slopes implied by the curve. However, the linear relationship still understates the actual relationships in the middle and overstates the relationship at the extremes (unless the independent variable has values only in a region where the curve is nearly linear). Figure 1.4 compares the S-shaped curve with the straight line; the gap between the two illustrates the nature of the error and the potential inaccuracy of linear regression.

Nonadditivity

The ceiling and floor create another conceptual problem besides nonlinearity in regression models of a dichotomous response. Regression typically assumes additivity—that the effect of one independent variable on the dependent variable stays the same regardless of the levels of the other independent variables. Models can include selected product terms to account for nonadditivity, but a binary dependent variable likely violates the additivity assumption for all combinations of the independent variables. If the value of one independent variable reaches a sufficiently high level to push the probability of the dependent variable to near 1 (or to near 0), then the effects of other variables cannot have much influence. Thus, the ceiling and floor make the influence of all the independent variables inherently nonadditive and interactive.

To return to the smoking example, those persons with 20 years of education have such a low probability of smoking that only a small difference exists between men and women; in other words, gender has little effect on smoking at high levels of education. In contrast, larger gender differences likely exist when education is lower and the probability of smoking is higher. Although the effect of gender on smoking likely varies with the level of education, additive regression models assume that the effect is identical for all levels of education (and the effect of education is identical for men and women). One can use interaction terms in a regression model to partly capture nonadditivity, but that does not address the nonadditivity inherent in all relationships in a probability model.

Problems of Statistical Inference

Even if a straight line approximates the nonlinear relationship in some instances, other problems emerge that, despite leaving the estimates unbiased, reduce their efficiency. The problems involve the fact that regression with a binary dependent variable violates the assumptions of normality and homoscedasticity. Both these problems stem from the existence of only two observed values for the dependent variable. Linear regression assumes that in the population, a normal distribution of error values around the predicted Y is associated with each X value, and that the dispersion of the error values for each X value is the same. The assumptions imply normal and similarly dispersed error distributions.

Yet, with a dummy variable, only two Y values and only two residuals exist for any single X value. For any value X_i, the predicted probability equals $b_0 + b_1X_i$. Therefore, the residuals take the value of

$$1 - (b_0 + b_1 X_i) \text{ when } Y_i \text{ equals } 1,$$

and

$$0 - (b_0 + b_1 X_i) \text{ when } Y_i \text{ equals } 0.$$

Even in the population, the distribution of errors for any X value will not be normal when the distribution has only two values.

The error term also violates the assumption of homoscedasticity or equal variances because the regression error term varies with the value of X. To illustrate this graphically, review Figure 1.1(b), which plots the relationship between X and a dummy dependent variable. Fitting a straight line that goes from the lower left to the upper right of the figure would define residuals as the vertical distance from the points to the line. Near the lower and upper extremes of *X,* where the line comes close to the floor of 0 and the ceiling of 1, the residuals are relatively small. Near the middle values of *X,* where the line falls halfway between the ceiling and floor, the residuals are relatively large. As a result, the variance of the errors is not constant (Greene, 2008, p. 775).[2]

While normality creates few problems with large samples, heteroscedasticity has more serious implications. The sample estimates of the population regression coefficients are unbiased, but they no longer have the smallest variance and the sample estimates of the standard errors will be too small. Thus, even with large samples, the standard errors in the presence of heteroscedasticity will be incorrect, and tests of significance will be biased in the direction of being too generous. Using robust standard errors or weighted least squares estimates can deal with this problem, but they do not solve the conceptual problems of nonlinearity and nonadditivity.

Transforming Probabilities Into Logits

To review, linear regression has problems in dealing with a dependent variable having only two values, a ceiling of 1 and a floor of 0: the same change in *X* has a different effect on *Y* depending on how close the curve corresponding to any *X* value comes to the maximum or minimum *Y* value. We need a transformation of the dependent variable that captures the decreasing effects of *X* on *Y* as the predicted *Y* value approaches the floor or ceiling. We need, in other words, to eliminate the floor and ceiling inherent in binary outcomes and probabilities.

The logistic function and logit transformation define one way to deal with the boundary problem. Although many nonlinear functions can represent the S-shaped curve (Agresti, 2013, Chapter 7), the logistic or logit transformation has become popular because of its desirable properties and relative simplicity. The logistic function takes probabilities as a nonlinear function of *X* in a way that represents the S-shaped curve in Figure 1.3(b). We will review the logistic function in more detail shortly. For now, simply note that the function defines a relationship between the values of *X* and the S-shaped curve in probabilities. As will become clear, the probabilities need to be transformed in a way that defines a linear rather than nonlinear relationship with *X*. The logit transformation does this.

Assume that each value of X_i has a probability of having a characteristic or experiencing an event, defined as P_i. Since the dependent variable has values of only 0 and 1, this P_i must be estimated, but it helps to treat the outcome in terms of probabilities for now. Given this probability, the logit transformation involves two steps. First, take the ratio of P_i to $1 - P_i$, or the odds of the outcome. Second, take the natural logarithm of the odds. The logit thus equals

$$L_i = \ln [P_i/(1 - P_i)],$$

or, in short, the logged odds.

It is worth seeing how the equation works with a few numbers. For example, if P_i equals .2, the odds equal .25 or .2/.8, and the logit equals –1.386, the natural log of the odds. If P_i equals .7, the odds equal 2.33 or .7/.3, and the logit equals 0.847. If P_i equals .9, the odds equal 9 or .9/.1, and the logit equals 2.197. Although the computational formula to convert probabilities into logits is straightforward, it requires some explanation to show its usefulness. It turns out to transform the S-shaped nonlinear relationship between independent variables and a distribution of probabilities into a linear relationship.

Meaning of Odds

The logit begins by transforming probabilities into odds. Probabilities vary between 0 and 1, and express the likelihood of an outcome as a proportion of both occurrences and nonoccurrences. Odds or $P/(1-P)$ express the likelihood of an occurrence relative to the likelihood of a nonoccurrence. Both probabilities and odds have a lower limit of 0, and both express the increasing likelihood of an outcome with increasing large positive numbers, but otherwise they differ.

Unlike a probability, odds have no upper bound or ceiling. As a probability gets closer to 1, the numerator of the odds becomes larger relative to the denominator, and the odds become an increasingly large number. The odds thus increase greatly when the probabilities change only slightly near their upper boundary of 1. For example, probabilities of .99, .999, .9999, .99999, and so on result in odds of 99, 999, 9999, 99999, and so on. Tiny changes in probabilities result in huge changes in the odds and show that the odds increase toward infinity as the probabilities come closer and closer to 1.

To illustrate the relationship between probabilities and odds, examine the values below:

P_i	.01	.1	.2	.3	.4	.5	.6	.7	.8	.9	.99
$1 - P_i$	.99	.9	.8	.7	.6	.5	.4	.3	.2	.1	.01
Odds	.01	.111	.25	.429	.667	1	1.5	2.33	4	9	99

Note that when the probability equals .5, the odds equal 1 or are even. As the probabilities increase toward one, the odds no longer have the ceiling of the probabilities. As the probabilities decrease toward 0, however, the odds still approach 0. At least at one end, then, the transformation allows values to extend linearly beyond the limit of 1.

Manipulating the formula for odds gives further insight into their relationship to probabilities. Beginning with the definition of odds (O_i) as the ratio of the probability to one minus the probability, we can with simple algebra express the probability in terms of odds:

$$P_i/(1 - P_i) = O_i \text{ implies that } P_i = \mathrm{O}_i/(1 + O_i).$$

The probability equals the odds divided by one plus the odds.[3]

Based on this formula, the odds can increase to infinity, but the probability can never equal or exceed one. No matter how large the odds become in the numerator, they will always be smaller by one than the denominator. Of course, as the odds become large, the gap between the odds and the odds plus 1 will become relatively small and the probability will approach (but not reach) one. To illustrate, the odds of 9 translate into a probability of .9, as 9/(9+1) = .9, the odds of 999 translate into a probability of .999 (999/1000 = .999), and the odds of 9999 translate into a probability of .9999, and so on.

Conversely, the probability can never fall below 0. As long as the odds equal or exceed 0, the probability must equal or exceed 0. The smaller the odds in the numerator become, the larger the relative size of the 1 in the denominator. The probability comes closer and closer to 0 as the odds come closer and closer to 0.

Usually, the odds are expressed as a single number, taken implicitly as a ratio to 1. Odds above 1 mean the outcome is more likely to occur than to not occur. Thus, odds of 10 imply the outcome will occur 10 times for each time it does not occur. Since the single number can be a fraction, there is no need to keep both the numerator or denominator as a whole number. The odds of 7 to 3 can be expressed equally well as a single number of 2.33 (to 1). Even odds equal 1 (1 occurrence to 1 nonoccurrence). Odds below 1 mean the outcome is less likely to occur than it is to not occur. If the probability equals .3, the odds are .3/.7 or .429. This means the outcome occurs .429 times per each time it does not occur. It could also be expressed as 42.9 occurrences per 100 nonoccurrences.

Comparing Odds

Expressed as a single number, any odds can be compared to another odds, only the comparison is based on multiplying rather than on adding. Odds of 9 to 1 are three times higher than odds of 3 to 1. Odds of 3 are one-third the size of odds of 9. Odds of .429 are .429 the size of even odds of 1, or half the size of odds of .858. In each example, one odds is expressed as a multiple of the other.

It is often useful to compare two different odds as a ratio. Consider the odds of an outcome for two different groups. The ratio of odds of 8 and 2 equals 4, which shows that the odds of the former group are four times (or 400%) larger than the latter group. If the odds ratio is below 1, then the odds of the first group are lower than the second group. An odds ratio of .5 means the odds of the first group are only half or 50% the size of the second group. The closer the odds ratio to 0, the lower the odds of the first group to the second. An odds ratio of one means the odds of both groups are identical. Finally, if the odds ratio is above one, the odds of the first group are higher than the second group. The greater the odds ratio, the higher the odds of the first group to the second.

To prevent confusion, keep in mind the distinction between odds and odds ratios. Odds refer to a ratio of probabilities, while odds ratios refer to ratios of odds (or a ratio of probability ratios). According to the 2016 GSS, for example, 65.9% of men and 57.2% of women favor legalization of marijuana. Since the odds of support for men equal 1.93 (.659/.341), it indicates that around 1.9 men support legalization for 1 who does not. The odds of support for legalization among women equal 1.34 (.572/.428) or about 1.3 women support legalization for 1 who does not. The ratio of odds of men to women equals 1.93/1.34 or 1.44. This odds ratio is a group comparison. It reflects the higher odds of supporting legalization for men than women. It means specifically that 1.44 men support legalization for each women who does.

In summary, reliance on odds rather than probabilities provides for meaningful interpretation of the likelihood of an outcome, and it eliminates the upper boundary. Odds will prove useful later for interpreting coefficients, but note now that creating odds represents the first step of the logit transformation.

Logged Odds

Taking the natural log of the odds eliminates the floor of 0 much as transforming probabilities into odds eliminates the ceiling of 1. Taking the natural log of:

odds above 0 but below 1 produce negative numbers;
odds equal to 1 produce 0; and
odds above 1 produce positive numbers.

(The logs of values equal to or below 0 do not exist; see the Appendix for an introduction to logarithms and their properties.)

The first property of the logit, then, is that, unlike a probability, it has no upper or lower boundary. The odds eliminate the upper boundary of probabilities, and the logged odds eliminate the lower boundary of probabilities as well. To see this, if $P_i = 1$, the logit is undefined because the odds of $1/(1 - 1)$ or $1/0$ do not exist. As the probability comes closer and closer to 1, however, the logit moves toward positive infinity. If $P_i = 0$, the logit is undefined because the odds equal zero $0/(1 - 0) = 0$ and log of 0 does not exist. As the probability comes closer and closer to 0, however, the logit proceeds toward negative infinity. Thus, the logits vary from negative infinity to positive infinity. The ceiling and floor of the probabilities (and the floor of the odds) disappear.

The second property is that the logit transformation is symmetric around the midpoint probability of .5. The logit when $P_i = .5$ is 0 ($.5/.5 = 1$, and the log of 1 equals 0). Probabilities below .5 result in negative logits because the odds fall below 1 and above 0; P_i is smaller than $1 - P_i$, thereby resulting in a fraction, and the log of a fraction results in a negative number (see the Appendix). Probabilities above .5 result in positive logits because the odds exceed 1 (P_i is larger than $1 - P_i$). Furthermore, probabilities the same distance above and below .5 (e.g., .6 and .4, .7 and .3, .8 and .2) have the same logits, but different signs (e.g., the logits for these probabilities equal, in order, .405 and –.405, .847 and –.847, 1.386 and –1.386). The distance of the logit from 0 reflects the distance of the probability from .5 (again noting, however, that the logits do not have boundaries as do the probabilities).

The third property is that the same change in probabilities translates into different changes in the logits. The principle is that as P_i comes closer to 0 and 1, the same change in the probability translates into a greater change in the logged odds. You can see this by example:

P_i	.1	.2	.3	.4	.5	.6	.7	.8	.9
$1 - P_i$	.9	.8	.7	.6	.5	.4	.3	.2	.1
Odds	.111	.25	.429	.667	1	1.5	2.33	4	9
Logit	–2.20	–1.39	–.847	–.405	0	.405	.847	1.39	2.20

A change in probabilities of .1 from .5 to .6 (or from .5 to .4) results in a change of .405 in the logit, whereas the same probability change of .1 from .8 to .9 (or from .2 to .1) results in a change of .810 in the logit. The change in the logit for the same change in the probability is twice as large at this extreme as in the middle. To repeat, the general principle is that small differences in probabilities result in increasingly larger differences in logits when the probabilities are near the bounds of 0 and 1.

Linearizing the Nonlinear

It helps to view the logit transformation as linearizing the inherent nonlinear relationship between X and the probability of Y. We would expect the same change in X to have a smaller impact on the probability of Y near the floor or ceiling than near the midpoint. Because the logit expands or stretches the probabilities of Y at extreme values relative to the values near the midpoint, the same change in X comes to have similar effects throughout the range of the logit transformation of the probability of Y. Without a floor or ceiling, in other words, the logit can relate linearly to changes in X. One can now treat a relationship between X and the logit transformation as linear. The logit transformation straightens out the nonlinear relationship between X and the original probabilities.

Conversely, the linear relationship between X and the logit implies a nonlinear relationship between X and the original probabilities. A unit change in the logit results in smaller differences in probabilities at high and low levels than at levels in the middle. Just as we translate probabilities into logits, we can translate logits into probabilities (the formula to do this is discussed shortly):

Logit	–3	–2	–1	0	1	2	3
P_i	.047	.119	.269	.5	.731	.881	.953
Change	—	.072	.150	.231	.231	.150	.072

A one-unit change in the logit translates into a greater change in probabilities near the midpoint than near the extremes. In other words, linearity in logits defines a theoretically meaningful nonlinear relationship with the probabilities.

Obtaining Probabilities From Logits

The logit transformation defines a linear relationships between the independent variables and a binary dependent variable. The linear relationship of X with the predicted logit appears in the following regression model:

$$\ln[P_i/(1 - P_i)] = b_0 + b_1X_i.$$

Like any linear equation, the coefficient b_0 shows the intercept or logged odds when X equals 0 and the b_1 coefficient shows the slope or the change in the logged odds for a unit change in X. The difference is that the dependent variable has been transformed from probabilities into logged odds.

To express the probabilities rather than the logit as a function of *X,* first take each side of the equation as an exponent. Because the exponent of a logarithm of a number equals the number itself (*e* of the ln *X* equals *X*), exponentiation or taking the exponential eliminates the logarithm on the left side of the equation:

$$P_i/1-P_i = e^{b_0+b_1X_i} = e^{b_0} \times e^{b_1X_i}.$$

Furthermore, the equation can be presented in multiplicative form because the exponential of $X + Y$ equals the exponential of *X* times the exponential of *Y.* Thus, the odds change as a function of the coefficients treated as exponents. Solving for P_i gives the following formula[4]:

$$P_i = (e^{b_0+b_1X_i})/(1+e^{b_0+b_1X_i}).$$

To simplify, define the predicted logit L_i as $\ln[P_i/(1-P_i)]$, which is equal to $b_0 + b_1X_i$. We can then replace the longer formula by L_i in the equation, remembering that L_i is the logged odds predicted by the value of X_i and the coefficients b_0 and b_1. Then

$$P_i = (e^{L_i})/(1+e^{L_i}).$$

This formula takes the probability as a ratio of the exponential of the logit to 1 plus the exponential of the logit. Given that e^{Li} produces odds, the formula corresponds to the equation $P_i = O_i/(1+O_i)$ presented earlier.

Moving from logits to exponents of logits to probabilities shows

L	–4.61	–2.30	–1.61	–.223	0	1.61	2.30	4.61	6.91
e^L	.01	.1	.2	.8	1	5	10	100	1000
$1+ e^L$	1.01	1.1	1.2	1.8	2	6	11	101	1001
P	.010	.091	.167	.444	.5	.833	.909	.990	.999

Note first that the exponentials of the negative logits fall between 0 and 1, and that the exponentials of the positive logits exceed 1. Note also that the ratio of the exponential to the exponential plus 1 will always fall below one—the denominator will always exceed the numerator by 1. The transformation of logits into probabilities replicates the S-shaped curve in Figures 1.3(b) and 1.4. With logits defining the *X*-axis and probabilities defining the *Y*-axis, the logits range from negative infinity to positive infinity, but the probabilities will stay within the bounds of 0 and 1.

Consider how this transformation demonstrates nonlinearity. For a one-unit change in *X, L* changes by a constant amount but *P* does not. The exponents in the formula for P_i make the relationship nonlinear. Consider an example. If $L_i = 2 + .3X_i$, the logged odds change by .3 for a one-unit change in *X* regardless of the level of *X*. If *X* changes from 1 to 2, *L* changes from 2 + .3 or 2.3 to 2 + .3 × 2 or 2.6. If *X* changes from 11 to 12, *L* changes from 5.3 to 5.6. In both cases, the change in *L* is identical. This defines linearity.

Take the same values of *X*, and the *L* values they give, and note the changes they imply in the probabilities:

X	1	2	11	12
L	2.3	2.6	5.3	5.6
e^L	9.97	13.46	200.3	270.4
$1 + e^L$	10.97	14.46	201.3	271.4
P	.909	.931	.995	.996
Change		.022		.001

The same change in *L* due to a unit change in *X* results in a greater change in the probabilities at lower levels of *X* and *P* than at higher levels. The same would show at the other end of the probability distribution.

This nonlinearity between the logit and the probability creates a fundamental problem of interpretation. We can summarize the effect of *X* on the logit simply in terms of a single linear coefficient, but we cannot do the same with the probabilities: the effect of *X* on the probability varies with the value of *X* and the level of the probability. The complications in interpreting the effects on probabilities require a separate chapter on the meaning of logistic regression coefficients. However, dealing with problems of interpretation proves easier having fully discussed the logic of the logit transformation.

One last note. For purposes of calculation, the formula for probabilities as a function of the independent variables and coefficients takes a somewhat simpler but less intuitive form:

$$P_i = (e^{b_0 + b_1 X_i})/(1 + e^{b_0 + b_1 X_i}),$$

$$P_i = 1/(1 + e^{-(b_0 + b_1 X_i)}),$$

$$P_i = 1/(1 + e^{-L_i}).$$

This gives the same result as the other formula.[5] If the logit equals –2.302, then we must solve for $P = e^{-2.302}/1+ e^{-2.302}$ or $1/1+ e^{-(-2.302)}$. The exponential of –2.302 equals approximately .1, and the exponential of the negative of –2.302 or 2.302 equals 9.994. Thus, the probability equals .1/1.1 or .091, or calculated alternatively equals 1/1 + 9.994 or .091. The same calculations can be done for any other logit value to get probabilities.

Summary

This chapter reviews how the logit transforms a dependent variable having inherent nonlinear relationships with a set of independent variables into a dependent variable having linear relationships with a set of independent variables. Logistic regression models (also called logit models) thus estimate the linear determinants of the logged odds or logit rather than the nonlinear determinants of the probabilities. Obtaining these estimates involves complexities left until later chapters. In the meantime, however, it helps to view logistic regression as analogous to linear regression on a dependent variable that has been transformed to remove the floor and ceiling.

Another justification of the logistic regression model and the logit transformation takes a different approach than offered in this chapter. It assumes that an underlying, unobserved, or latent continuous dependent variable exists. It then derives the logistic regression model by making assumptions about the shape of the distribution of the underlying unobserved values and its relationship to the observed values of 0 and 1 for the dependent variable. This derivation ends up with the same logistic regression model but offers some insights that may be useful. See, for example, Long (1997, pp. 40–51), Maddala and Lahiri (2009, p. 333), or Greene (2008, pp. 776–777).[6]

In linearizing the nonlinear relationships, logistic regression also shifts the interpretation of coefficients from changes in probabilities to less intuitive changes in logged odds. The loss of interpretability with the logistic coefficients, however, is balanced by the gain in parsimony: the linear relationship with the logged odds can be summarized with a single coefficient, while the nonlinear relationship with the probabilities is less easily summarized. Efforts to interpret logistic regression coefficients in meaningful and intuitive ways define the topic of the next chapter.