*To Seth*

# LOGISTIC REGRESSION: A PRIMER

**FRED C. PAMPEL**
*University of Colorado, Boulder*

## 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, have their income fall below the poverty level, vote for one candidate, favor or oppose an issue, 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. 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 Dummy Dependent Variable

A binary qualitative 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 $X$.

To give a simple example, the 1994 General Social Survey (GSS) of the National Opinion Research Corporation asked respondents if they smoke. Assigning those who smoke a score of 1 and those who do not a score of 0 creates a dichotomous dependent variable. Taking smoking ($S$) as a function of years of completed education ($E$) and a dummy variable for gender ($G$) with females coded 1 produces the regression equation:

$$S = .661 - .029 * E + .004 * G.$$

The coefficient for education indicates that for a 1-year increase in education, the probability of smoking goes down by .029, the proportion smoking goes down by .029, or the percent smoking goes down by 2.9. Male respondents with no education have a predicted probability of smoking of .661 (the intercept). A male with 10 years of education has a predicted probability of smoking of .371 ($.661 - .029 * 10$). One could also say that the model predicts 37% of such respondents smoke. The dummy variable coefficient shows females have a probability of smoking .004 higher than for males. With no education, women have a predicted probability of smoking of .665 ($.661 + .004$).

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. Together, the problems prove serious enough to require use of an alternative to ordinary regression with qualitative dependent variables.

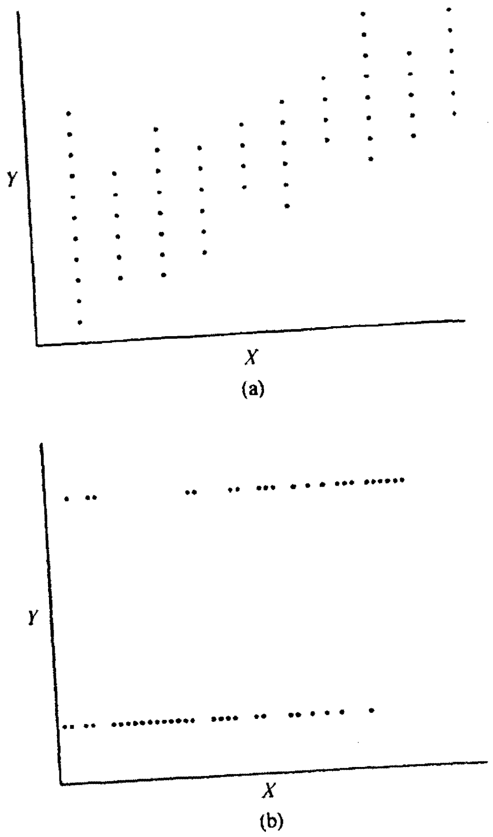
### *Problems of Functional Form*

The conceptual problem with linear regression with a dichotomous 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 can extend upward toward positive infinity as the values of the independent variables increase indefinitely, and extend downward toward negative infinity as the values of the independent variables decrease indefinitely. 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 zero) 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 reveal the density of cases along the two lines. Jittering reduces overlap of the scatterplot points by adding random variation to each case. In Figure 1.2, the jittered 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.

The risk of predicted probabilities below 0 or above 1 can, depending also on the range of values of the independent variable, increase with the skew of the dichotomous dependent variable. With a split of around 50:50, predicted values tend to fall toward the center of the probability distribution. In the previous example of smoking (where the split equals 28:72), the lowest predicted value of .081 occurs for



(a)

(b)

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

males with the maximum education of 20; the highest predicted value of .665 occurs for females with the minimum education of 0. A more skewed dependent variable from the GSS asks respondents if they are a member of any group that aims to protect or preserve the environment. With the 10% saying yes coded 1 and others coded 0, a regression on education and gender gives

$$B = -.024 + .008 * E - .006 * G.$$

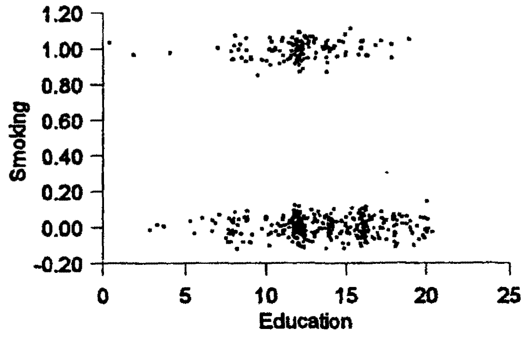


Figure 1.2. Jittered scatterplot for a binary dependent variable, smoking or nonsmoking by years of education.

The intercept shows the nonsensical probability that a male with no education will have a predicted probability of belonging below zero. Although a problem in general, reliance on the assumption of linearity in this particular model proves particularly inappropriate.[1]

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 $X$ 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 1.3(a)).

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 indefinitely, the nonlinear curve bends 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 experiencing an event from .95 to .96 requires a larger change in $X$ than it does to produce

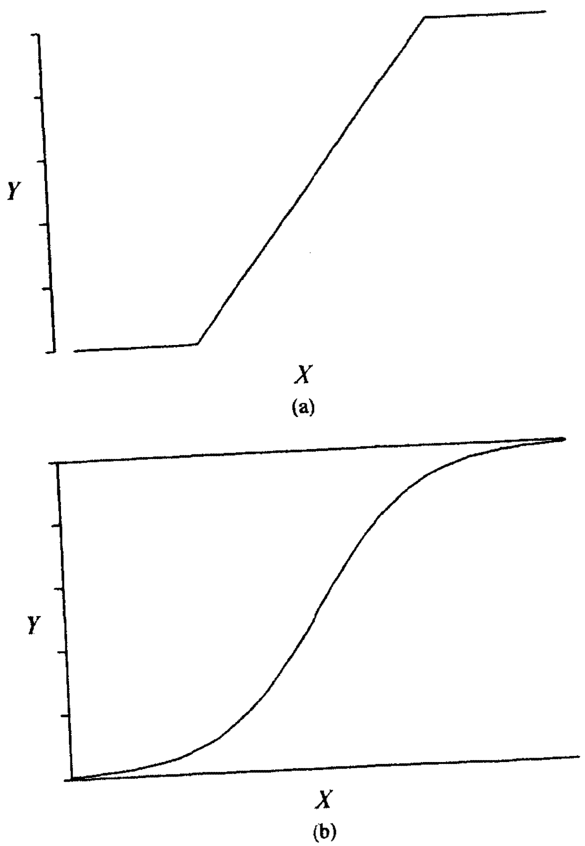


Figure 1.3. (a) Truncated linear relationship, (b) S-shaped curve.

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 $40,000 to $50,000 would increase that likelihood more than an increase from $200,000 to $210,000. 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 has 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 teens and early twenties than at younger or older ages. Few will marry under age 15 despite growing a year older, and few unmarried by 50 will likely marry by age 51. However, the change from age 21 to 22 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 1.3(b), where the curve levels off and approaches the ceiling of 1 and 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 a 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.

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 variable on the dependent variable stays the same regardless of the levels of

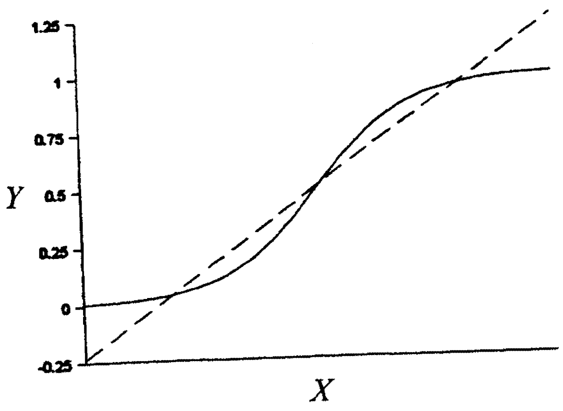


Figure 1.4. Linear versus curvilinear relationship.

the other independent variables. Models can include selected product terms to account for nonadditivity, but a dichotomous 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 can exist between men and women; in other words, sex can have little effect on smoking at high levels of education. In contrast, larger sex differences likely exist when education is lower and the probability of smoking is higher. Although the effect of sex on smoking likely varies with the level of education, additive regression models incorrectly assume that the effect of sex on smoking is identical for all levels of education (and the effect of education is identical for both sexes).

### *Problems of Statistical Inference*

Even if a straight line approximates the nonlinear relationship in some instances, some problems emerge that, despite leaving the estimates unbiased, reduce their efficiency. The problems involve the fact that regression with a dummy 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

$$1 - (b_0 + b_1X_i) \text{ when } Y_i \text{ equals } 1,$$

and

$$0 - (b_0 + b_1X_i) \text{ when } Y_i \text{ equals } 0.$$

Even in the population, the distribution of errors for any $X$ value cannot 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 $X$.[2] To illustrate this graphically, review Figure 1.1(b), which plots the relationship between $X$ and a dichotomous 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.

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 are biased. Thus, even with large samples, the standard errors in the presence of heteroscedasticity will be incorrect, and tests of significance will be invalid. Technical means of weighing least squares estimates can deal with this problem, but more importantly do not solve the conceptual problems of nonlinearity and nonadditivity. Use of regression with a dummy dependent variable consequently remains inappropriate.

## Transforming Probabilities into Logits

Linear regression faces a problem in dealing with a dependent variable with a ceiling and a floor: 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 to allow for 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 probabilities.

Although many nonlinear functions can represent the S-shaped curve, the logistic or logit transformation, because of its desirable properties and relative simplicity, has become popular. To illustrate the logit transformation, assume that each case 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 experiencing the event. 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.

For example, if $P_i$ equals .2 for the first case, its odds equals .25 or .2/.8, and its logit equals −1.386, the natural log of the odds. If $P_i$ for the second case equals .7, its odds equal 2.33 or .7/.3, and its logit equals 0.847. If $P_i$ equals .9 for the third case, its odds equals 9 or .9/.1, and its logit equals 2.197. Although the computational formula to transform probabilities into logits is straightforward, it requires some explanation to show its usefulness. It turns out to describe the relationship between independent variables and a distribution of probabilities defined by a dichotomous dependent variable.

### *Meaning of Odds*

The logit begins by transforming probabilities into odds. Probabilities vary between 0 and 1, and express the likelihood of an event as a proportion of both occurrences and nonoccurrences. Odds express the likelihood of an occurrence relative to the likelihood of a nonoccurrence. Both probabilities and odds have a lower limit of zero, and both express the increasing likelihood of an event 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

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_i$ | .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 zero, however, the odds still approach zero. At least at one end, then, the transformation allows values to extend linearly beyond the previous limit.

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:

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

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

Based on this formula, 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. Conversely, the probability can never fall below 0. As long as the odds equal or exceed 0, the probability must equal or exceed zero. 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 zero 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. Thus, odds of 10 imply an event 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). Thus, even odds equal 1 (1 occurrence to 1 nonoccurrence). Odds below 1 mean the event 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 event occurs .429 times per each time it does not occur. It could also be expressed as 42.9 occurrences per 100 nonoccurrences.

Expressed as a single number, any odds can be compared to another odds. Odds of 9 to 1 are three times higher than odds of 3. 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. 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 for 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 zero, 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 1994 GSS, for example, 29.5% of men and 13.1% of women own a gun, Since the odds of gun ownership for men equal .418 (.295: .705), it indicates that around 4 men own a gun for 10 who do not. The odds of gun ownership for women equal .151 or about 1.5 women own a gun for 10 who do not. The ratio of odds of men to women equal .418: .151 or 2.77, which means that the odds of gun ownership are nearly three times higher for men than women.

In summary, reliance on odds rather than probabilities provides for meaningful interpretation of the likelihood of events, but eliminates the upper boundary. Odds will prove useful later in 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 produces negative numbers;
odds equal to 1 produces 0; and
odds above 1 produces positive numbers.

(The logs of values equal to or below zero 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 bound of probabilities as well. To see this, if $P_i = 1$, the logit is undefined because the odds of 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 log of the odds of 0/1 or 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 problem of a ceiling and floor in the probabilities (or a floor in odds) disappears.

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 one ($P_i$ is larger than $1 - P_i$). Further, 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 the probabilities listed above 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 simple 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 compute a linear relationship between $X$ and the logit transformation. 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 linear relationships between the independent variables and the logit dependent variable imply nonlinear relationships with probabilities. The linear relationship of $X$ to the predicted logit appears in

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

To express the probabilities rather than the logit as a function of $X$, first take each side of the equation as an exponent. Since the logarithm of a number as an exponent 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:

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

Further, 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 formula[4]:

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

Since the logit $L_i$ equals $b_0 + b_1X_i$, we can 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

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

This formula takes the probability as a ratio of the exponential of the logit to 1 plus the exponential of the logit. Given that $e^{L_i}$ 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 one. 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. However, as the exponential gets larger, the difference between the numerator and the denominator declines (in other words, the extra one unit in the denominator becomes increasingly small relative to the other value in the numerator). Further, the ratio can never fall below zero since the exponentials of both negative and positive numbers end up positive and since the ratio of two positive numbers always ends up positive. Given the boundaries of the probabilities, the example shows that the larger $L$, the larger $e^L$, and the larger $P$.

This transformation also 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 $P_i$ makes 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. |

Hence, 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 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.

### *An Alternative Formula*

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:

$$P_i = e^{b_0+b_1X_i}/\left(1+e^{b_0+b_1X_i}\right),$$
$$P_i = 1/\left(1+e^{-(b_0+b_1X_i)}\right),$$
$$P_i = 1/\left(1+e^{-L_i}\right).$$

In this formula, you need to take the exponential after taking the negative of the logit. The probability then equals 1 divided by 1 plus the exponential of the negative of the logit. This gives exactly the same result as the other formula.[5]

Either formula works to translate logits into probabilities. 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.[6] Logistic regression models (sometimes also called logit models) thus estimate the linear determinants of the logged odds or logit rather than the nonlinear determinants of probabilities. Obtaining these estimates involves complexities left until later chapters. In the meantime, however, it helps to view logistic regression in simple terms as regression on a dependent variable that transforms nonlinear relationships into linear relationships.

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, but the nonlinear relationship with the probabilities cannot be so simply summarized. Efforts to interpret logistic regression coefficients in a meaningful, yet relatively simple way define the topic of the next chapter.

## 2. INTERPRETING LOGISTIC REGRESSION COEFFICIENTS

Although it simplifies the estimation issues to come, treating logistic regression as a form of regression on a dependent variable transformed into logged odds helps describe the underlying logic of the procedure. However, as is true for nonlinear transformations more generally, the effects of the independent variables in logistic regression have multiple interpretations. Effects exist for probabilities, odds, and logged odds, and the interpretations of each effect have both advantages and disadvantages.

To preview, the effects of the independent variables on the logged odds are linear and additive—each $X$ variable has the same effect on the logged odds regardless of its level or the level of other $X$ variables—but the units of the dependent variable, logged odds, have little intuitive meaning. The effects of the independent variables on the probabilities have intuitive meaning, but are nonlinear and nonadditive—each $X$ variable has a different effect on the probability depending on its level and the level of the other independent variables. Despite the interpretable units, the effects on probabilities cannot be simply summarized in the form of a single coefficient. The interpretation of the effects of the independent variables on the odds offers a compromise between the previous alternatives. The odds have more intuitive appeal than the logged odds, and can express effects in single coefficients. The effects on odds are multiplicative rather than additive, but still have a straightforward interpretation. Other ways to interpret the effects of the independent variables exist. The ratios of the coefficients to their standard errors obviously have importance in interpreting sample results. Also, various attempts to standardize the coefficients for the independent variables and compare their relative size may prove helpful.

This chapter examines each of these ways to interpret effects in logistic regression. Further, it examines the variations in each interpretation for continuous and dummy independent variables.

### Logged Odds

The first interpretation directly uses the coefficients obtained from the estimates of the logistic regression. The logistic regression coefficients simply show the change in the predicted logged odds of experiencing an event or having a characteristic for a one-unit change in the independent variables. The coefficients have exactly the same interpretation as the coefficients in regression except that the units of the dependent variable represent the logged odds. For example, Browne (1997, p. 246) uses logistic regression to predict participation in the labor force of 922 female heads of household between ages 18 and 54 in 1989. The logistic regression coefficient of .13 for years employed shows that each additional year of employment increases the logged odds of current participation in the labor force by .13.

For dummy variables, a change in one unit implicitly compares the indicator group to the reference or omitted group. Browne uses dummy variables for high school dropouts and high school graduates to compare their labor force participation to those women with some college education (the reference group). The coefficients of −1.29 and −.68 for these two dummy variables indicate that the logged odds of being in the labor force are 1.29 lower for high school dropouts than for those with some college, and are .68 lower for high school graduates than for those with some college. Excepting the metric of the dependent variable, this interpretation represents nothing different from that used for dummy variables in ordinary regression.

These coefficients represent the relationship, as in ordinary regression, with a single coefficient. Regardless of the value of $X$—small, medium, or large—or the values of the other independent variables, a one-unit change has the same effect on the dependent variable. According to the model, the difference in the logged odds of participation between a woman with 1 year of experience and a woman with 2 years of experience equals the difference in the logged odds of participation between a woman with 21 years of experience and a woman with 22 years of experience. Similarly, the effect of years employed in the model does not differ between high school dropouts, high school graduates, and those with some college. All one needs to do is copy the coefficient from the printout. Indeed, logistic regression aims to simplify the nonlinear and nonadditive relationships inherent in treating probabilities as dependent variables.

Note also that logistic regression, as in linear regression, can include product terms to represent interactive relationships and polynomial terms to represent curvilinear relationships. The product and squared terms in logistic regression have much the same interpretation as in linear regression, except that the units of the dependent variable take the form of logged odds. Logistic regression already contains nonadditivity and nonlinearity in the relationships between the independent variables and probabilities, but can further model nonadditivity and nonlinearity in the relationship between the independent variables and the logged odds (DeMaris, 1992).

Despite the simplicity of their interpretation, the logistic regression coefficients, as mentioned, lack a meaningful metric. Statements about the effects of variables on changes in logged odds reveal little about the relationships and do little to help explain the substantive results. Researchers need means to interpret the substantive meaning or importance of the coefficients other than merely reporting the expected changes in logged odds.

### Odds

The second interpretation comes from transforming the logistic regression coefficients so that the independent variables affect the odds rather than the logged odds of the dependent variable. To find the effects on the odds, simply take the exponent or antilogarithm of the logistic regression coefficients. As in the two variable model that follows, exponentiating both sides of the logistic regression equation eliminates the log of the odds and shows the influences of the variables on the odds,

$$
\begin{aligned}
\ln(P/1-P) &= b_0 + b_1X_1 + b_2X_2,\\
e^{\ln(P/1-P)} &= e^{b_0+b_1X_1+b_2X_2},\\
P/1-P &= e^{b_0} * e^{b_1X_1} * e^{b_2X_2}.
\end{aligned}
$$

As noted in the last chapter, the antilog of the log of a value equals the value itself, and the left side of the equation equals the odds. In addition, since the exponent of $(X + Y)$ equals the exponent of $X$ times the exponent of $Y$, the right-hand side of the equation becomes multiplicative rather than additive.

The odds are a function of the exponentiated constant ($e^{b_0}$) multiplied by the exponentiated product of the coefficient and $X_1(e^{b_1X_1})$ and the exponentiated product of the coefficient and $X_2(e^{b_2X_2})$. In simple terms, the effect of each variable on the odds (rather than the logged odds) comes from taking the antilog of the coefficients. If not already presented in the computer output, the exponentiated coefficients can be obtained from most any calculator by typing the coefficient and then the $e^x$ function. The exponentiated coefficients of .13, −1.29, and −.68 from Browne's study of women's labor force participation equal 1.14, .28, and .51.

The fact that the equation determining the odds is multiplicative rather than additive affects the interpretation of the exponentiated coefficients. In an additive equation, a variable has no effect when its coefficient equals 0. The predicted value of the dependent variable sums the values of the variables times the coefficients; when adding 0, the predicted value does not change. In a multiplicative equation, the predicted value of the dependent variable does not change when multiplied by a coefficient of 1. Therefore, 0 in the additive equation corresponds to 1 in the multiplicative equation. Further, the exponential of a positive number exceeds 1 and the exponential of a negative number falls below 1 (but above zero, as the exponential of any number is always greater than zero).

For the exponentiated coefficients, then, a coefficient of 1 leaves the odds unchanged, a coefficient greater than 1 increases the odds, and a coefficient smaller than 1 decreases the odds. Moreover, the more distant the coefficient from 1 in either direction, the greater the effect in changing the odds. For example, the exponentiated coefficient for years of employment, 1.14, indicates that a 1-year increase in employment multiplies the odds of labor force participation by 1.14 or increases the odds by a factor of 1.14. If the odds of participation for someone employed 12 years equals 4.88, the odds of participation for someone employed 13 years equals 4.88 * 1.14 or 5.56. The odds of participation for someone employed 14 years in turn equals 5.56 * 1.14 or 6.34.[7]

In terms of odds ratios, dividing the odds of someone with 13 years of experience by the odds of someone with 12 years of experience gives the exponentiated logistic regression coefficient: 5.56/4.48 = 1.14. Thus, the coefficient shows the ratio of odds for a one-unit increase in the independent variable.

For dummy variables, a similar interpretation follows. The exponentiated coefficient for the high school dropout dummy variable, .28, indicates that a one-unit increase in the variable multiplies the odds of labor force participation by .28. Of course, a one-unit increase compares high school dropouts to the reference group of those with some college. In either case, multiplying by .28 substantially lowers the odds. If the odds of participation for those with some college equal 15.6, the odds of participation for high school dropouts equal 15.6 * .28 or 4.37. For high school graduates, the exponentiated coefficient of .51 indicates that the odds of participation are .51 times smaller than for those with some college. Their odds would equal 15.6 * .51 or 7.96. In terms of odds ratios, the exponentiated coefficient for the dummy variable equals the ratio of odds for the dummy variable group to the odds for the reference group.

Since the distance of an exponentiated coefficient from 1 indicates the size of the effect, a simple calculation can further aid in interpretation. The difference of a coefficient from 1 exhibits the increase or decrease in the odds for a unit change in the independent variable. In terms of a formula, the exponentiated coefficient minus 1 and times 100 gives the percentage increase or decrease due to a one-unit change in the independent variable:

$$\%\Delta = (e^b - 1) * 100.$$

For years of employment, the exponentiated coefficient says that the odds of participating in the labor force increase by 14% for an increase of 1 year of employment experience. This appears more meaningful than to say the logged odds increase by .13.[8] The size of the effect on the odds also depends on the units of measurement of the independent variables—the change in odds for variables measured in different units do not warrant direct comparison. Still, the interpretation of percentage change in the odds has intuitive appeal.

Turning to the dummy variables, the percentage change of the exponentiated logistic regression coefficient for high school dropouts equals (.28 − 1) * 100 or −72. This means that the odds of participating are 72% lower for high school dropouts than for those with some college. The exponentiated coefficient for high school graduates of .51 indicates that their odds of participating are 49% lower than for those with some college.

In interpreting the exponentiated coefficients, remember that they refer to multiplicative changes in the odds rather than probabilities. It is easy to say that an additional year of work experience makes participation 1.14 times more probable or otherwise imply probabilities rather than odds (DeMaris, 1995, p. 1960). More precisely, the odds of participation are 1.14 times as large or 14% larger for an additional year of work.

## Probabilities

The third strategy of interpreting the logistic regression coefficients involves translating the effects on logged odds or odds into the effects on probabilities. Since the relationships between the independent variables and probabilities are nonlinear and nonadditive, they cannot be fully represented by a single coefficient. The effect on the probabilities has to be identified at a particular value or set of values. The choice of values to use in evaluating the effect of variables on the probabilities depends on the concerns of the researcher and the nature of the data, but an initial strategy has the advantage of simplicity: examine the effect on the probability for a typical case.

### *Continuous Independent Variables*

One quick way to gauge the influence of a continuous variable on probabilities involves calculating the linear slope of the tangent of the nonlinear curve at any single point. The slope of the tangent line is defined by the partial derivative of the nonlinear equation relating the independent variables to the probabilities, but more intuitively represents a straight line that meets the logistic curve at a single point without crossing to the other side of the curve. Figure 2.1 depicts the tangent line where the logistic curve intersects $Y = P = .76$. The tangent line identifies the slope only at that particular point, but allows for easy interpretation. Its slope shows the linear change in the probability for a one-unit change in the independent variable defined at a single point on the logistic curve.

The change in probability or the linear slope of the tangent line comes from a simple equation for the partial derivative. The partial derivative reveals the change in the probability for an infinitely small change in $X$, but also defines the slope of the tangent line or the change in the tangent line due to a one-unit change in $X$ at that value (as discussed shortly, it does not equal the actual change in the logistic regression curve due to a one-unit change in $X$). The partial derivative, also referred to as the marginal or instantaneous effect, equals

$$\partial P/\partial X_k = b_k * P * (1 - P).$$

Simply multiply the logistic regression coefficient by the selected probability $P$ and 1 minus the probability.

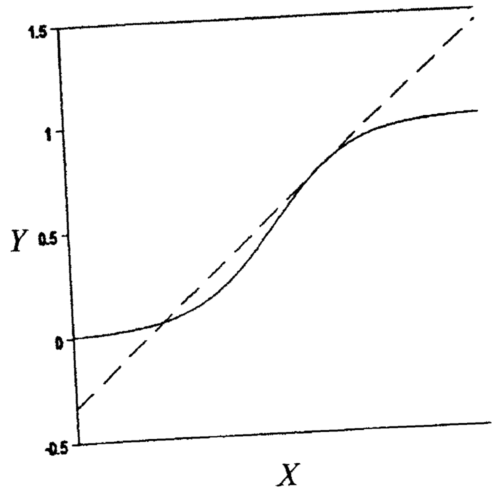


Figure 2.1. Tangent line of logistic curve at $Y = P = .76$.

The formula for the partial derivative nicely reveals the nonlinear effects of an independent variable on probabilities. The effect of $b$ (in terms of logged odds) translates into a different effect on the probabilities depending on the level of $P$. The effect will be at its maximum when $P$ equals .5 since $.5 * .5 = .25$, $.6 * .4 = .24$, $.7 * .3 = .21$ and so on. The closer $P$ comes to the ceiling or floor, the smaller the value $P * (1 - P)$, and the smaller the effect a unit change in $X$ has on the probability.

Multiplying the coefficient times $.5 * .5$ shows the maximum effect on the probabilities, but may overstate the influence for a sample in which the split on the dependent variable is not so even. Substituting the mean of the dependent variable, $P$, in the formula gives a more typical effect. In Browne's example, the logistic regression coefficient for years employed equals .13; the mean of the dependent variable, the expected probability of participating, equals .83; and the probability of not participating equals .17. Multiplying all three gives a value of .018. An increase of 1 year of employment increases the probability of participation by .018 or almost 2% at the mean. The effect reaches its maximum of .032 when $P = .5$.

As an alternative to the mean, we might compute the predicted probability for a typical case on the independent variables, and use that probability to calculate the partial derivative. Substituting the means of the continuous variables and the value of the modal category for dummy variables into the logistic regression equation yields the predicted logged odds for that case. Transforming the predicted logged odds into a predicted probability allows calculation of the effects on probabilities for that case.

In much the same way, a researcher might compute a predicted probability for a range of values on the independent variables and present the marginal effects for the extremes as well as the middle of the sample (Long, 1997, p. 64). Allowing all the other variables to take their mean values, calculate the predicted probabilities when one variable takes values −2, −1, 0, 1, and 2 standard deviations from the mean. Then use these probabilities to calculate marginal effects. Alternatively, calculate probabilities and the associated marginal effects when the independent variable takes its maximum, mean, and minimum values. Long (1997) discusses a number of others ways—including the use of both tables and graphs—to present a more complete summary of the range of influences of a variable on probabilities.

The formula for the partial derivative demonstrates the nonadditive as well as the nonlinear nature of the relationships with probabilities: the effect of one independent variable on the probabilities varies with $P$, and $P$ varies with the values of other independent variables. When $X_2$ is at its mean, it might predict $P$ near .5 and $X_1$ would have a relatively large marginal effect. When $X_2$ is near its maximum, it might predict a $P$ near 1 and $X_1$ would have a relatively small marginal effect. The effect of $X_1$ on the probabilities, in other words, varies with the values of other independent variables and predicted $P$ values. This means that the independent variables interact in determining probabilities (remember that the effects of the variables on the logged odds are linear and additive).

The inherent nonlinear and nonadditive influence of the determinants on probabilities limits the value of any single summary coefficient. Given the difficulties of describing a nonlinear and nonadditive relationship with a single coefficient, analysts disagree over whether it is valuable to even calculate a single partial derivative (DeMaris, 1990, 1993; Roncek, 1993). Critics of the procedure view the resulting coefficient as misleading, and little better than using linear regression. Even so, the tendency of researchers to think in terms of proportions or probabilities may warrant use of the slope of the tangent at the mean of the dependent variable or other points on the logistic curve as a supplement to other interpretations.

### *Dummy Independent Variables*

The partial derivative works best with continuous variables for which small changes in the independent variables that define the tangent have meaning. For dummy variables, the relevant change occurs from 0 to 1, and the tangent line for small changes in $X$ makes less sense. Instead, it is possible to compute predicted probabilities for each group, and then subtract the two probabilities to measure the group differences in probabilities. The partial derivative of the coefficient for a dummy variable may approximate the group difference in probabilities, but calculating the predicted probabilities gives the exact difference. Remember, however, that the calculated group difference in probabilities, like the partial derivative, varies with the point on the logistic curve, the $X$ values, and the $P$ values.

To make the calculation, select a starting probability from which to evaluate the effect of the dummy variable. With this value serving as the probability for the omitted group, calculate the predicted probability for the dummy variable group. Subtracting these two probabilities shows the difference in the probability between the two groups evaluated at the selected starting point (Peterson, 1985). The mean of the dependent variable may serve as the probability of the omitted group, but other values of special interest may work equally well as the starting point. Choosing other $P$ values for the omitted group, although appropriate and useful, will produce different results.

More precisely, follow these steps. (a) Find the logged odds of $P$ or the predicted logit for the omitted group. (b) To get the predicted logit for the dummy variable group, add the logistic regression coefficient to the predicted logit for the omitted group. (c) Compute the probability from the predicted logit for the dummy variable group using the formula listed below and in Chapter 1. (d) Subtract $P$ from the probability for the dummy variable group to obtain the between-group difference in probabilities (or the effect of the dummy variable on probabilities).

In formula, the steps take the form,

$$L_o = \ln(P_o/(1 - P_o)) \text{ logit for the omitted group,}$$
$$L_d = L_o + b_d \text{ logit for the dummy variable group,}$$
$$P_d = 1/1 + e^{-L_d} \text{ probability for the dummy variable group,}$$
$$P_d - P_o \text{ difference in probabilities.}$$

In Browne's example, using the mean of the dependent variable or .83 as $P_o$ and the $b$ for high school dropouts of −1.29 (with women with some college serving as the omitted group), follow the previous steps,

$$L_o = \ln(P_o/(1 - P_o)) = \ln(.83/.17) = 1.586 \text{ logit for women with some college,}$$