# 5

# LOGIT REGRESSION

Logistic regression, more commonly called *logit regression*, is used when the response variable is dichotomous (i.e., binary or 0–1). The predictor variables may be quantitative, categorical, or a mixture of the two.

Sometimes ordinary bivariate or multiple regression is used in this situation. When this is done, the model is called the *linear probability model*. The linear probability model provides a useful introduction to the logit regression model, so we consider it first.

## 5.1. THE LINEAR PROBABILITY MODEL

Suppose that we wish to analyze the effect of education on current contraceptive use among fecund, nonpregnant, currently married women aged 35–44. Our variables are

$C$: contraceptive use (1 if using, 0 otherwise)
$E$: number of completed years of education

The scatterplot of $C$ against $E$, for a few representative points, might look something like the plot in Figure. 5.1. An ordinary bivariate regression line can be fitted through the points by ordinary least squares (OLS), as shown. The estimated equation of the line has the

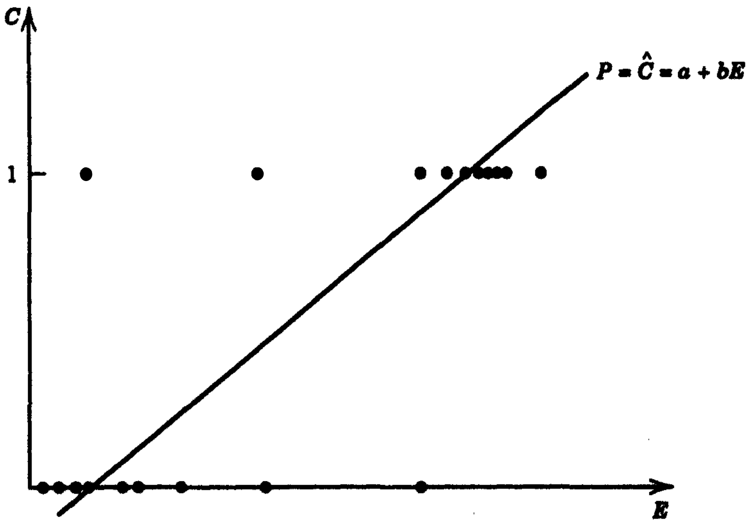


**Figure 5.1.** The linear probability model: probability of contraceptive use predicted by education. **Notes:** $C$ denotes contraceptive use (1 if using, 0 otherwise), and $E$ denotes number of completed years of education. $\hat{C}$, denoting the value of $C$ predicted from the regression, is interpreted as the probability of currently using contraception.

form

$$\hat{C} = a + bE \tag{5.1}$$

where $\hat{C}$ is the value of $C$ predicted by the regression.

The observed value of $C$ can assume only two values, 0 and 1. In contrast, the value of $C$ predicted by the regression line, $\hat{C}$, can assume a continuum of values. For most observed values of the predictor variable, $\hat{C}$ will have a value between 0 and 1. *We interpret $\hat{C}$ as the probability that a woman with a specified level of education is currently using contraception.*

Because $\hat{C}$ is interpreted as a probability, (5.1) may also be written as

$$P = a + bE \tag{5.2}$$

where $P$ denotes the estimated probability of use. If so desired, (5.2) can be elaborated by adding more predictor variables, interaction terms, quadratic terms, and so on.

Although the linear probability model has the advantage of simplicity, it suffers from some serious disadvantages:

1. *The estimated probability $P$ can assume impossible values.* At the lower left end of the line in Figure 5.1, $P$ is negative, and at the upper right end, $P$ exceeds unity.
2. *The linearity assumption is seriously violated.* According to this assumption, the expected value of $C$ at any given value of $E$ falls on the regression line. But this is not possible for the parts of the line for which $P < 0$ or $P > 1$. In these regions, the observed points are either all above the line or all below the line.
3. *The homoscedasticity assumption is seriously violated.* The variances of the $C$ values tend not to behave properly either. The variance of $C$ tends to be much higher in the middle range of $E$ than at the two extremes, where the values of $C$ are either mostly zeros or mostly ones. In this situation, the equal-variance assumption is untenable.
4. *Because the linearity and homoscedasticity assumptions are seriously violated, the usual procedures for hypothesis testing are invalid.*
5. *$R^2$ tends to be very low.* The fit of the line tends to be very poor. Because the response variable can assume only two values, 0 and 1, the $C$ values tend not to cluster closely about the regression line.

For these reasons, the linear probability model is seldom used, especially now that alternative models such as logit regression are widely available in statistical software packages such as SAS, LIMDEP, BMDP, and SPSS.

## 5.2. THE LOGIT REGRESSION MODEL

Instead of a straight line, it seems preferable to fit some kind of sigmoid curve to the observed points. *By a sigmoid curve, we mean a curve that resembles an elongated S or inverted S laid on its side.* The tails of the sigmoid curve level off before reaching $P = 0$ or $P = 1$, so that the problem of impossible values of $P$ is avoided.

Two hypothetical examples of sigmoid curves are shown in Figure 5.2. In Figure 5.2a, the response variable is $C$, as before, and the curve increases from near zero at low levels of education to near unity

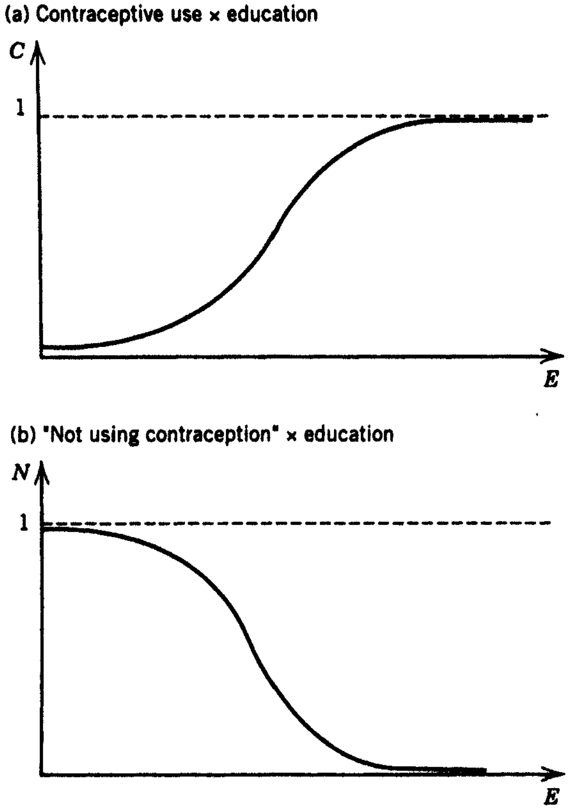


**Figure 5.2.** Two examples of sigmoid curves.

at high levels of education. In Figure 5.2b, the response variable is "not using contraception," denoted by $N$, where $N$ is one if the woman is not using contraception and zero otherwise. Now the curve decreases from near unity at low levels of education to near zero at high levels of education.

The sigmoid curve assumes that the predictor variable has its largest effect on $P$ when $P$ equals 0.5, and that the effect becomes smaller in absolute magnitude as $P$ approaches 0 or 1. (The effect of $E$ on $P$ is measured by the slope of a line tangent to the curve at each specified value of $E$.) In many situations, this is a reasonably accurate portrayal of reality.

### 5.2.1. The Logistic Function

What mathematical form should we assign to the sigmoid curve? Although there are many possibilities, the logistic function tends to be preferred, partly because it is comparatively easy to work with mathematically, and partly because it leads to a model (the logit regression model) that is comparatively easy to interpret.

There is, however, a certain amount of arbitrariness in the choice of functional form to represent the sigmoid curve, and results from the model depend to some extent on which functional form is chosen. Another functional form that is frequently used is the cumulative normal distribution, which forms the basis of the probit regression model. The logit and probit models usually yield similar but not identical results. We do not consider the probit model any further in this book.

The basic form of the logistic function is

$$\boxed{P = \frac{1}{1 + e^{-Z}}} \tag{5.3}$$

where $Z$ is the predictor variable and $e$ is the base of the natural logarithm, equal to 2.71828. . . . Throughout this chapter we shall view (5.3) as an estimated model, so that $P$ is an estimated probability.

If numerator and denominator of the right side of (5.3) are multiplied by $e^Z$, the logistic function in (5.3) can be written alternatively as

$$P = \frac{e^Z}{1 + e^Z} = \frac{\exp(Z)}{1 + \exp(Z)} \tag{5.4}$$

where $\exp(Z)$ is another way of writing $e^Z$. Equation (5.3), or equivalently (5.4), is graphed in Figure 5.3a.

A property of the logistic function, as specified by (5.3), is that when $Z$ becomes infinitely negative, $e^{-Z}$ becomes infinitely large, so that $P$ approaches 0. When $Z$ becomes infinitely positive, $e^{-Z}$ becomes infinitesimally small, so that $P$ approaches unity. When $Z = 0$, $e^{-Z} = 1$, so that $P = .5$. Thus the logistic curve in Figure 5.3a has its "center" at $(Z, P) = (0, .5)$.

To the left of the point $(0, .5)$, the slope of the curve (i.e., the slope of a line tangent to the curve) increases as $Z$ increases. To the right of this point, the slope of the curve decreases as $Z$ increases. *A point with this property is called an inflection point.*

Suppose that we have a theory that $Z$ is a cause of $P$. Then the slope of the curve at a particular value of $Z$ measures the effect of $Z$ on $P$ at that particular value of $Z$. *Therefore, to the left of the*

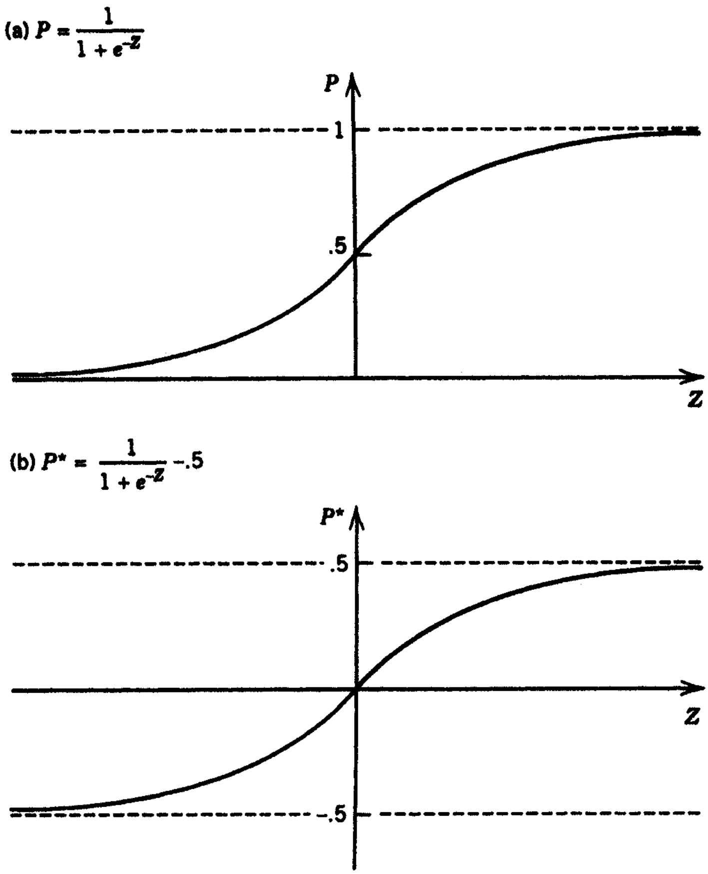


**Figure 5.3.** The logistic function.

*inflection point, the effect of Z on P increases as Z increases. To the right of the inflection point, the effect of Z on P decreases as Z increases.* The effect of $Z$ on $P$ attains its maximum at the inflection point. Effects are not constant over the range of the predictor variable, as they are in the simple bivariate regression model.

Another property of the logistic curve is that it is symmetric about its inflection point, as can be demonstrated as follows: First, subtract 0.5 from $P$, so that the curve is moved downward by 0.5, as shown in Figure 5.3b. The inflection point is now at the origin instead of (0, .5). Except for this downward translation, the shape of the curve is the same as before. Of course, when we do this translation, $P$ can no longer be interpreted as a probability, because $P$ now ranges between $-0.5$ and $+0.5$. We therefore rename $P$ as $P^*$. The equation of the curve is now

$$P^*(Z) = \frac{1}{1 + e^{-Z}} - .5 \tag{5.5}$$

where we write $P^*$ as $P^*(Z)$ to emphasize the functional dependence of $P^*$ on $Z$.

The symmetry of the curve, as graphed in Figure 5.3b, can now be demonstrated by showing that the vertical upward distance from the horizontal axis to the curve at any given positive value of $Z$ equals the vertical downward distance from the horizontal axis to the curve at $-Z$—in other words, by proving that $P^*(Z) = -P^*(-Z)$. This is done by showing that the equation

$$\frac{1}{1 + e^{-Z}} - .5 = -\left[\frac{1}{1 + e^{Z}} - .5\right] \tag{5.6}$$

reduces to an identity, an exercise that is left up to the reader. [An identity is an equation that is true regardless of the value of the unknown—in this case $Z$. The validity of (5.6) can be demonstrated by reducing (5.6) to the form $0 = 0$.] A less rigorous demonstration that the formula is an identity can be accomplished by substituting in a few randomly chosen values of $Z$ and checking to see that equality between the left and right sides of (5.6) is preserved.

### 5.2.2. The Multivariate Logistic Function

Equation (5.3) is bivariate. How can we make it multivariate? Suppose that $Z$, instead of being a single predictor variable, is a linear function of a set of predictor variables:

$$Z \equiv b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{5.7}$$

(Note that $Z$ is not a response variable in this equation.) This expression can be substituted for $Z$ in the formula for the logistic function in (5.3):

$$\boxed{P = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k)}}} \tag{5.8}$$

All the basic properties of the logistic function are preserved when this substitution is done. The function still ranges between 0 and 1 and achieves its maximum rate of change, with respect to change in any of the $X_i$, at $P = .5$.

As a simple example, suppose that (5.7) assumes the very simple form $Z = -X$. Then (5.8) becomes

$$P = \frac{1}{1 + e^{X}} \tag{5.9}$$

*When $Z = -X$, as in equation* (5.9), *the graph of $P$ against $X$ is a reversed sigmoid curve, which is* 1 *at* $-\infty$ *and* 0 *at* $+\infty$.

As a slightly more complicated example, suppose that (5.7) takes the form $Z = a + bX$, where $a$ and $b$ are parameters that are fitted to the data:

$$P = \frac{1}{1 + e^{-(a+bX)}} \tag{5.10}$$

Equation (5.10) can be rewritten as

$$P = \frac{1}{1 + e^{-b(a/b+X)}} \tag{5.11}$$

from which it is evident that the curve is centered at $X = -a/b$ instead of $X = 0$. *The constant term $a/b$ shifts the curve to the left or right, depending on whether $a/b$ is positive or negative, and the coefficient $b$ stretches or compresses the curve along the horizontal dimension, depending on whether $|b|$ (the absolute value of $b$) is less than or greater than 1. If $b$ is negative, the curve goes from 1 to 0 instead of 0 to 1 as $X$ increases.* (The reader may demonstrate these properties by picking some alternative sets of values of $a$ and $b$ and then constructing, for each set, a graph of $P$ against $X$ for (5.3) with $a + bX$ substituted for $Z$.) The introduction of the parameters $a$ and $b$ makes the model more flexible, so that a better fit can be achieved.

### 5.2.3. The Odds and the Logit of $P$

The logit of $P$ is derived from the logistic function

$$P = \frac{1}{1 + e^{-Z}} \qquad \text{(5.3 repeated)}$$

From (5.3) it follows that

$$1 - P = 1 - \frac{1}{1 + e^{-Z}} = \frac{e^{-Z}}{1 + e^{-Z}} \tag{5.12}$$

Dividing (5.3) by (5.12) yields

$$\frac{P}{1 - P} = e^{Z} \tag{5.13}$$

Taking the natural logarithm (base $e$) of both sides of (5.13), we obtain

$$\boxed{\log \frac{P}{1 - P} = Z} \tag{5.14}$$

The quantity $P/(1 - P)$ is called the *odds*, denoted more concisely as $\Omega$ (uppercase omega), and the quantity $\log[P/(1 - P)]$ is called the *log odds* or the *logit of P*. Thus

$$\boxed{\text{Odds} \equiv \frac{P}{1 - P} \equiv \Omega} \tag{5.15}$$

and

$$\boxed{\text{logit } P \equiv \log \frac{P}{1 - P} \equiv \log \Omega} \tag{5.16}$$

The definition of the odds in (5.15) corresponds to everyday usage. For example, one speaks of the odds of winning a gamble on a horse race as, say, "75 : 25", meaning .75/(1 − .75) or, equivalently, 75/(100 − 75). Alternatively, one speaks of "three-to-one" odds, which is the same as 75 : 25.

With these definitions, and with the expression in (5.7) substituted for $Z$, (5.14) can be written alternatively as

$$\boxed{\text{logit } P = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k} \tag{5.17}$$

$$\boxed{\log \frac{P}{1 - P} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k} \tag{5.18}$$

or

$$\boxed{\log \Omega = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k} \qquad (5.19)$$

Equations (5.17)–(5.19) are in the familiar form of an ordinary multiple regression equation. This is advantageous, because some of the statistical tools previously developed for multiple regression can now be applied to logit regression.

### 5.2.4. Logit Regression Coefficients as Measures of Effect on Logit *P*

The discussion of effects is facilitated by considering a simple specification of (5.17) that can be fitted to data from the 1974 Fiji Fertility Survey. Suppose that we wish to investigate the effect of education, residence, and ethnicity on contraceptive use among fecund, nonpregnant, currently married women aged 35–44. Sample size is $n = 954$. Our variables are

$P$: estimated probability of contraceptive use
$E$: number of completed years of education
$U$: 1 if urban, 0 otherwise
$I$: 1 if Indian, 0 otherwise

For $U$, "otherwise" means rural, and for $I$, "otherwise" means Fijian, as in Chapter 3.

Our model is

$$\text{logit } P = a + bE + cU + dI \qquad (5.20)$$

which, when fitted to the Fiji data, is

$$\text{logit } P = \underset{(.181)}{-\ .611} + \underset{(.025)}{.055E} + \underset{(.153)}{.378U} + \underset{(.166)}{1.161I} \qquad (5.21)$$

Numbers in parentheses under the coefficients are standard errors. In every case, the coefficient is at least twice its standard error, indicating that the coefficient differs significantly from zero at the 5 percent level with a two-tailed test. We shall consider later how the fitting is done.

Just as in ordinary multiple regression, the coefficients of the predictor variables can be interpreted as effects if there is theoretical justification for doing so. Because (5.21) is in the form of a multiple regression equation, we can immediately say, for example, that the effect of a one-year increase in education, controlling for residence and ethnicity, is to increase logit $P$ by .055. Similarly, the effect of being urban, relative to rural, controlling for education and ethnicity, is to increase logit $P$ by .378. And the effect of being Indian, relative to Fijian, controlling for education and residence, is to increase logit $P$ by 1.161.

*The trouble is that logit P is not a familiar quantity, so that the meanings of these effects are not very clear.*

To some extent, we can clarify the interpretation of these effects by writing logit $P$ as $\log[P/(1-P)]$ and then examining the relationship between $P, P/(1-P)$, and $\log[P/(1-P)]$. We observe first that $P/(1-P)$ is a monotonically increasing function of $P$, and $\log[P/(1-P)]$ is a monotonically increasing function of $P/(1-P)$, as shown in Figure 5.4. [A function $f(P)$ is monotonically increasing if an increase in $P$ always generates an increase in $f(P)$.] Figure 5.4 also shows the graph of $\log[P/(1-P)]$ against $P$.

Monotonicity means that if $P$ increases, $P/(1-P)$ and $\log[P/(1-P)]$ also increase; if $P$ decreases, $P/(1-P)$ and $\log[P/(1-P)]$ also decrease. Conversely, if $\log[P/(1-P)]$ increases, $P/(1-P)$ and $P$ also increase; if $\log[P/(1-P)]$ decreases, $P/(1-P)$ and $P$ also decrease.

In (5.21), a one-year increase in $E$ generates an increase of .055 in $\log[P/(1-P)]$, which, because of monotonicity, goes hand in hand with increases in $P/(1-P)$ and $P$. *In other words, if b is positive, the effects of E on* $\log[P/(1-P)]$, $P/(1-P)$, *and P are positive. If b is negative, the effects of E on* $\log[P/(1-P)]$, $P/(1-P)$, *and P are all negative*. Similar statements can be made about the effects of $U$ and $I$.

This is helpful, but still not very satisfactory. The next section further sharpens the conceptualization of "effect."

### 5.2.5. Odds Ratios as Measures of Effect on the Odds

From (5.20) and (5.16), our estimated model may be expressed as

$$\log \Omega = a + bE + cU + dI \qquad (5.22)$$

Taking the exponential of both sides [i.e., taking each side of (5.22) as

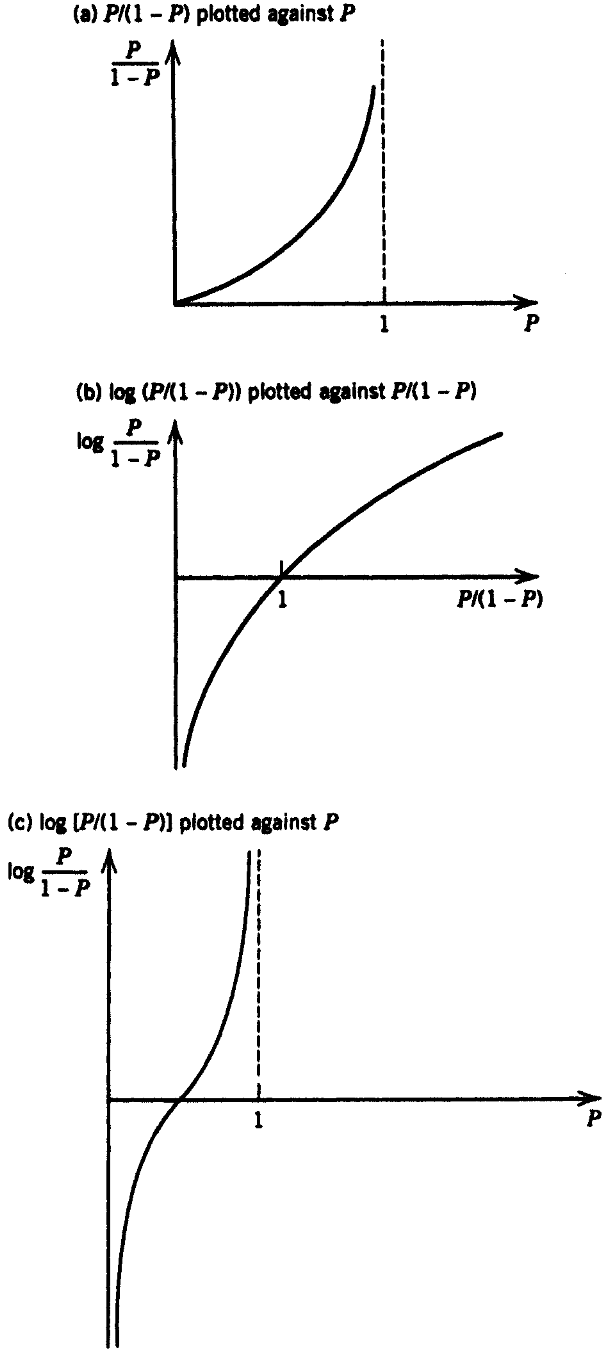


**Figure 5.4.** Graphic demonstration that $P/(1-P)$ and $\log(P/(1-P))$ are both monotonically increasing functions of $P$.

a power of $e$], we obtain

$$\Omega = e^{a+bE+cU+dI} \tag{5.23}$$

Suppose we increase $E$ by one unit, holding $U$ and $I$ constant. Denoting the new value of $\Omega$ as $\Omega^*$, we have

$$\begin{aligned}\Omega^* &= e^{a+b(E+1)+cU+dI} \\ &= e^{a+bE+cU+dI+b} \\ &= e^{a+bE+cU+dI}e^b \\ &= \Omega e^b\end{aligned} \tag{5.24}$$

which can be written alternatively as

$$\frac{\Omega^*}{\Omega} = e^b \tag{5.25}$$

From (5.24), it is evident that a one-unit increase in $E$, holding other predictor variables constant, multiplies the odds by the factor $e^b$. The quantity $e^b$ is called an *odds ratio*, for reasons that are obvious from (5.25).

Equation (5.25) can be arrived at more directly by noting that, because (5.22) is in the form of a multiple regression equation, the effect of a one-unit increase in $E$, holding other predictor variables constant, is to increase $\log \Omega$ by $b$ units. Thus

$$\log \Omega^* = \log \Omega + b \tag{5.26}$$

Taking each side of (5.26) as a power of $e$, we obtain

$$\Omega^* = \Omega e^b \tag{5.27}$$

which yields (5.25) when both sides are divided by $\Omega$.

Let us review what has happened. *The original coefficient $b$ represents the additive effect of a one-unit change in $E$ on the log odds of using contraception. Equivalently, the odds ratio $e^b$ represents the multiplicative effect of a one-unit change in $E$ on the odds of using contraception.* Insofar as the odds is a more intuitively meaningful concept than the log odds, $e^b$ is more readily understandable than $b$ as a measure of effect.

The above discussion indicates that the logit model may be thought of as either an additive model or a multiplicative model, depending on how the response variable is conceptualized. ***When we consider the log odds as the response variable, the logit model is an additive model, as in ordinary multiple regression. But when we consider the odds as the response variable, the logit model is a multiplicative model***, regarding the definition and interpretation of effects.

So far we have considered the effect of a one-unit change in $E$, which is a quantitative variable. What about a change in $U$, the dummy variable indicating urban or rural residence? If we consider the effect of a one-unit change in $U$ (from 0 to 1) on the odds of using contraception, holding $E$ and $I$ constant, then, starting from (5.23) and using the same logic as before, we obtain

$$\Omega^* = \Omega e^c \tag{5.28}$$

In other words, the effect of being urban, relative to rural, controlling for education and ethnicity, is to multiply the odds of using contraception by $e^c$. Similarly, the effect of being Indian, relative to Fijian, controlling for education and residence, is to multiply the odds by $e^d$.

In our Fijian example [see equation (5.21) above], we have that $e^b = e^{.055} = 1.057$, $e^c = e^{.378} = 1.459$, and $e^d = e^{1.161} = 3.193$. Therefore, a one-year increase in $E$, holding $U$ and $I$ constant, multiplies the odds by 1.057 (equivalent to a 5.7 percent increase). A one-unit increase in $U$ (from 0 to 1—i.e., from rural to urban), holding $E$ and $I$ constant, multiplies the odds by 1.459 (a 45.9 percent increase). A one-unit increase in $I$ (from 0 to 1—i.e., from Fijian to Indian), holding $E$ and $U$ constant, multiplies the odds by 3.193 (a 219.3 percent increase).

### 5.2.6. The Effect on the Odds When the Predictor Variable Is Categorical with More Than Two Categories

Let us alter our model of the effect of education, residence, and ethnicity on contraceptive use, as follows:

$P$: estimated probability of contraceptive use
$M$: 1 if medium education, 0 otherwise
$H$: 1 if high education, 0 otherwise
$U$: 1 if urban, 0 otherwise
$I$: 1 if Indian, 0 otherwise

This is the same model as before, except that education is redefined as a categorical variable with three categories: low, medium, and high, with low as the reference category.

In log odds form, the model is

$$\log \Omega = a + bM + cH + dU + fI \tag{5.29}$$

From (5.29), we can calculate the following values of $\log \Omega$ for low, medium, and high education by setting $(M, H)$ alternatively to (0, 0), (1, 0), and (0, 1):

| | | |
|---|---|---|
| Low education | $(H = 0, M = 0)$: | $\log \Omega_L = a + dU + fI$ |
| Medium education | $(H = 0, M = 1)$: | $\log \Omega_M = a + b + dU + fI$ |
| High education | $(H = 1, M = 0)$: | $\log \Omega_H = a + c + dU + fI$ |

From these values of high, medium, and low education, we obtain

$$\begin{aligned} \log \Omega_M - \log \Omega_L &= (a + b + dU + fI) - (a + dU + fI) \\ &= b \end{aligned} \tag{5.30}$$

$$\begin{aligned} \log \Omega_H - \log \Omega_L &= (a + c + dU + fI) - (a + dU + fI) \\ &= c \end{aligned} \tag{5.31}$$

$$\begin{aligned} \log \Omega_M - \log \Omega_H &= (a + b + dU + fI) - (a + c + dU + fI) \\ &= b - c \end{aligned} \tag{5.32}$$

Equations (5.30)–(5.32) can be rewritten

$$\log \Omega_M = \log \Omega_L + b \tag{5.33}$$

$$\log \Omega_H = \log \Omega_L + c \tag{5.34}$$

$$\log \Omega_M = \log \Omega_H + (b - c) \tag{5.35}$$

Equation (5.33) says that the effect of medium education, relative to low education, controlling for $U$ and $I$, is to increase $\log \Omega$ by $b$. Equation (5.34) says that the effect of high education, relative to low education, controlling for $U$ and $I$, is to increase $\log \Omega$ by $c$. Equation (5.35) says that the effect of medium education, relative to high education, controlling for $U$ and $I$ is to increase $\log \Omega$ by $b - c$. It doesn't matter what values $U$ and $I$ are held constant at, because terms in $U$ and $I$ cancel in (5.30)–(5.32). Except for the form of the response variable, these results are the same as those obtained in Chapter 2 for ordinary multiple regression.

If equations (5.33)–(5.35) are taken as powers of $e$, these equations become

$$\Omega_M = \Omega_L e^b \tag{5.36}$$

$$\Omega_H = \Omega_L e^c \tag{5.37}$$

$$\Omega_M = \Omega_H e^{b-c} \tag{5.38}$$

Equation (5.36) says that the effect of medium education, relative to low education, controlling for $U$ and $I$, is to multiply the odds by $e^b$. Equation (5.37) says that the effect of high education, relative to low education, controlling for $U$ and $I$, is to multiply the odds by $e^c$. Equation (5.38) says that the effect of medium education, relative to high education, controlling for $U$ and $I$, is to multiply the odds by $e^{b-c}$.

In equations (5.33)–(5.35), the effects on log $\Omega$, which are additive, are $b$, $c$, and $b - c$. In equations (5.36)–(5.38), the effects on $\Omega$, which are multiplicative, are $e^b$, $e^c$, and $e^{b-c}$.

Because (5.29) is in the form of a multiple regression equation, we can immediately say that the effect on logit $P$ of, say, medium education, relative to high education, does not depend on which education category is chosen as the reference category. It follows that the effects on $P/(1 - P)$ and $P$ itself do not depend on which category of education is chosen as the reference category. Changing the reference category does not require one to rerun the model. One can derive the education coefficients in the new model from the education coefficients in the original model. (See Section 2.4.2 of Chapter 2.) The coefficients of the other predictor variables are unaffected by changing the reference category for education.

### 5.2.7. The Effect of the Predictor Variables on the Risk $P$ Itself

Let us return to our example of the effects of education ($E$), residence ($U = 1$ if urban, 0 otherwise), and ethnicity ($I = 1$ if Indian, 0 otherwise) on contraceptive use, where education is once again conceptualized as a quantitative variable, namely, the number of completed years of education.

If we wish to look at the effects of the predictor variables directly on the risk $P$, we must go back to the basic form of the logistic function:

$$P = \frac{1}{1 + e^{-(a+bE+cU+dI)}} \tag{5.39}$$

What is the effect on $P$ of a one-unit increase in $E$, holding $U$ and $I$ constant? Denoting the new value of $P$ as $P^*$, we get

$$P^* = \frac{1}{1 + e^{-(a + b(E+1) + cU + dI)}} \tag{5.40}$$

Both the difference, $P^* - P$ (representing the additive effect), and the ratio, $P^*/P$ (representing the multiplicative effect), are functions not only of $b$ but also of $a$, $c$, $d$, $E$, $U$, and $I$. Terms in $a$, $c$, $d$, $E$, $U$, and $I$ do not cancel as they do in the case of the log odds (where all that remains in the difference is $b$) or the odds (where all the remains in the ratio is $e^b$). Therefore, both the additive effect and the multiplicative effect of a one-unit change in $E$ on $P$ depend on the levels of $E$, $U$, and $I$.

One way to handle this problem is to set $U$ and $I$ equal to their means in the sample and then to calculate and tabulate values of $P$ for $E = 0, 1, 2, \ldots$. We shall return to this approach when we discuss how to present logit regression results in a multiple classification analysis (MCA) format.

Because the effects of the predictor variables on $P$ are not as simple as the effects on $\Omega$ and $\log \Omega$, logit regression results are often presented as effects on $\Omega$ or $\log \Omega$.

### 5.2.8. Interactions

Let us now complicate the model considered in Section 5.2.7 by adding an interaction between residence and ethnicity. For reasons that will become clear shortly, it is convenient to represent logit $P$ as $\log \Omega$. Our estimated model is then

$$\log \Omega = a + bE + cU + dI + fUI \tag{5.41}$$

Because (5.41) is in the form of an ordinary multiple regression equation, we can say immediately that the effect of a one-unit increase in $U$ (from 0 to 1) on $\log \Omega$, holding $E$ and $I$ constant, is to change $\log \Omega$ by $c + fI$ (see Section 2.6 in Chapter 2). Denoting the new value of $\Omega$ as $\Omega^*$, we therefore have that

$$\log \Omega^* = \log \Omega + (c + fI) \tag{5.42}$$

If we take each side of (5.42) as a power of $e$, (5.42) becomes

$$\Omega^* = \Omega e^{c+fI} \tag{5.43}$$

If $f$ is positive, the multiplicative effect of a one-unit increase in $U$ increases as $I$ increases from 0 to 1 (i.e., from Fijian to Indian). If $f$ is negative, the multiplicative effect of a one-unit increase in $U$ decreases as $I$ increases from 0 to 1.

It can similarly be shown that a one-unit increase in $I$, holding $E$ and $U$ constant, increases $\log \Omega$ by $d + fU$ and multiplies $\Omega$ by $e^{d+fU}$. If $f$ is positive, the multiplicative effect of a one-unit increase in $I$ increases as $U$ increases from 0 to 1 (i.e., from rural to urban). If $f$ is negative, the multiplicative effect of a one-unit increase in $I$ decreases as $U$ increases from 0 to 1.

*Note that, in the context of* (5.43), *interaction between U and I means that the multiplicative effect of U on* $\Omega$ *depends on the level of I. Similarly, the multiplicative effect of I on* $\Omega$ *depends on the level of U.*

### 5.2.9. Nonlinearities

Let us further complicate the model by adding a nonlinearity to the effect that $E$ has on $\log \Omega$:

$$\log \Omega = a + bE + cE^2 + dU + fI + gUI \tag{5.44}$$

Again, because (5.44) is in the form of a multiple regression equation, we can say immediately that the (instantaneous) effect of a one-unit increase in $E$, holding $U$ and $I$ constant, is to change $\log \Omega$ by $b + 2cE$ (see Section 2.7.1 in Chapter 2). Denoting the new value of $\Omega$ as $\Omega^*$, we therefore have that

$$\log \Omega^* = \log \Omega + (b + 2cE) \tag{5.45}$$

Taking each side of (5.45) as a power of $e$, we can rewrite (5.45) as

$$\Omega^* = \Omega e^{b+2cE} \tag{5.46}$$

If $c > 0$, the multiplicative effect of a one-unit increase in $E$ on $\Omega$ increases as $E$ increases. If $c < 0$, the multiplicative effect of a one-unit increase in $E$ on $\Omega$ decreases as $E$ increases.

*Note that, in the context of* (5.46), *nonlinearity in E means that the multiplicative effect of E on* $\Omega$ *depends on the level of E.*

## 5.3. STATISTICAL INFERENCE

### 5.3.1. Tests of Coefficients

The computer output for logit regression includes estimated standard errors and $p$ values for the coefficients. The standard error of a coefficient, $b$, is interpreted as the standard error of the sampling distribution of $b$, just as in ordinary regression. As before in this book, we shall accept the standard errors on faith and not be concerned about the details of the statistical theory underlying their derivation.

From a coefficient $b$ and its standard error $s_b$, a $Z$ *value* can be calculated as

$$\boxed{Z \equiv \frac{b - \beta}{s_b}} \tag{5.47}$$

where $\beta$ is the hypothesized value of the coefficient in the underlying population. Usually our null hypothesis is that $\beta = 0$. then (5.47) reduces to

$$Z \equiv \frac{b}{s_b} \tag{5.48}$$

If our hypothesized value of the coefficient in the underlying population is correct, then, for large samples, the sampling distribution of $Z$ is approximately standard normal with mean 0 and variance 1. How large is large? A rule of thumb that is sometimes used is the following: As long as the ratio of sample size to the number of parameters to be estimated is at least 20, the normal approximation is usually satisfactory. The sampling distribution of $Z$ for small samples is more complicated and is not considered in this book.

Given the coefficients and their standard errors and $Z$ values, hypothesis testing and the construction of confidence intervals are done in the same way as in ordinary multiple regression, using the table of standard normal probabilities (Table B2, Appendix B). *One does not use the table of t critical points in Table B.1 (Appendix B).* In Chapter 2 on multiple regression, the formulae analogous to (5.47) and (5.48) had $t$ in place of $Z$. In logit regression, however, the sampling distribution of $(b - \beta)/s_b$ does not conform to Student's $t$ distribution.

In logit regression as in ordinary multiple regression, one can optionally print out a regression coefficient covariance matrix and test whether two coefficients (e.g., the coefficients of $M$ and $H$, denoting medium and high education) differ significantly from each other. The test procedure is the same as that described earlier in Section 2.9.2 of Chapter 2.

If one is dealing with the odds instead of the log odds, then one needs to construct confidence intervals for $e^b$ instead of $b$. If the boundaries of the confidence interval for $b$ are $u$ and $v$, then the boundaries for the confidence interval for $e^b$ are calculated as $e^u$ and $e^v$.

### 5.3.2. Testing the Difference Between Two Models

As we shall discuss in more detail later, the *likelihood* $L$ is the probability of observing our particular sample data under the assumption that the fitted model is true. Thus $L$ is somewhere between 0 and 1. The *log likelihood, or* $\log L$ (log to the base $e$), is often mathematically more convenient to work with than $L$ itself. *Because $L$ lies between* 0 *and* 1, $\log L$ *is negative.*

Suppose that we have two logit regression models which have the same response variable but different sets of predictor variables, where the second model has all the predictor variables included in the first, plus at least one more. We say that the first model is *nested* in the second model.

Let us denote the likelihood of the first model by $L_1$, and the likelihood of the second model by $L_2$. One way to test whether the two models differ significantly from each other might be to test whether $L_1$ differs significantly from $L_2$, or, equivalently, whether $L_2 - L_1$ differs significantly from zero. But his is not possible, because the sampling distribution of $L_2 - L_1$ is not known.

Alternatively one might think of testing whether $L_1/L_2$ differs significantly from 1. But the sampling distribution of $L_1/L_2$ is not known either.

Fortunately, a statistic for which the sampling distribution is known is $-\log(L_1/L_2)^2$, where $L_1 < L_2$. This can be written as

$$\begin{aligned} -\log(L_1/L_2)^2 &= -2\log(L_1/L_2) \\ &= -2(\log L_1 - \log L_2) \\ &= (-2\log L_1) - (-2\log L_2) \qquad (5.49) \end{aligned}$$

The quantity $L_1/L_2$ is called the *likelihood ratio*. The condition $L_1 < L_2$, which is not restrictive, simply means that the first model is nested in the second model. When $L_1 < L_2, -\log(L_1/L_2)^2$ is positive.

*We can use* (5.49) *to test whether the second model fits the data significantly better than the first model. The test is a simple* $\chi^2$ *test. It turns out that* $-\log(L_1/L_2)^2$ *(i.e., the positive difference in* $-2\log L$ *between the two models) is distributed approximately as* $\chi^2$ *with degrees of freedom equal to the difference in the number of coefficients to be estimated in the two models.* The word "distributed" refers here to the sampling distribution of the difference in $-2\log L$; if we took repeated samples from the same underlying population, each time fitting the two models and computing the difference in $-2\log L$ between the two fitted models, we would find that the difference in $-2\log L$ would be distributed as just mentioned.

For example, suppose as a real-life example, for fecund, nonpregnant, currently married women aged 35–44 in the 1974 Fiji Fertility Survey, that the two fitted models are

$$\text{logit } P = .363 - .034E + .597U \tag{5.50}$$

and

$$\text{logit } P = -.611 + .055E + .378U + 1.161I \tag{5.51}$$

where $P$ denotes the probability of using contraception, $E$ denotes number of completed years of education, $U$ denotes residence (1 if urban, 0 otherwise), and $I$ denotes ethnicity (1 if Indian, 0 otherwise).

The $-2\log L$ statistic is 1264.6 for the first model and 1213.1 for the second model. (These values are automatically printed out by the SAS program for binary logit regression.) The difference in $-2\log L$ between the two models is therefore 51.5. The difference in the number of coefficients to be estimated is 1, so d.f. = 1. Consulting the $\chi^2$ table in Appendix B (Table B.4), we find that $\chi^2$ with d.f. = 1 is 10.8 for $p = .001$, which is the highest level of significance included in the table. Because our observed value of $\chi^2$ is 51.5, which is greater than the critical point of 10.8, we can say that the two models differ with an observed level of significance of $p < .001$. We conclude that the model in (5.51) fits the data much better than the model in (5.50).

In (5.50) and (5.51), $-2\log L$ is approximately 1200, which means that $\log L$ is approximately $-600$ and $L$ is approximately $e^{-600}$, which is approximately $10^{-261}$. This extremely small number, representing the likelihood of the observed data under the assumption that the model is true, might lead one to think that the model fits poorly. It must be remembered, however, that each individual observation has a probability, and that the overall likelihood is the product of these individual probabilities. Each individual probability is a number less than one. If the sample size is large, the product of the individual probabilities will be an extremely small number, even if the individual probabilities are fairly close to one. Therefore, a very small likelihood does not necessarily mean a poor fit.

Very frequently, the two models compared are the model in which we are interested, which we may refer to as the *test model*, and what is often called the *intercept model*. For example, if the test model is specified by (5.51), the corresponding intercept model is

$$\text{logit } P = a \tag{5.52}$$

where $a$ is a constant to be fitted. In effect, we ask whether the test model differs significantly from a baseline model with no predictor variables. Computer programs for logit regression typically print out not only the value of $-2 \log L$ for both the test model and the intercept model, but also the $p$ value for the difference in $-2 \log L$ between these two models. In this case, there is no need to consult a $\chi^2$ table, because the computer does it for us. For the Fiji data, the logit regression output for the test model in (5.51) indicates a $p$ value of .000 for the difference between this model and the intercept model. Because .000 is a rounded value, this means that $p < .0005$. In other words, the test model fits the data much better than the intercept model.

## 5.4. GOODNESS OF FIT

In multiple regression, the traditional indicator of goodness of fit is $R^2$, which measures the proportion of variation in the response variable that is explained by the predictor variables. In the case of logit regression, one could also calculate the proportion of variation in the response variable that is explained by the predictor variables, but in this case it is impossible for the observed values of the response variable, which are either 0 or 1, to conform exactly to the fitted values of $P$. The maximum value of this proportion depends on the mean and variance of $P$ (Morrison, 1972).

An alternative measure of goodness of fit may be derived from the likelihood statistic. Let $L_0$ denote the likelihood for the fitted intercept model, and let $L_1$ denote the likelihood for the fitted test model. Define pseudo-$R^2$ as

$$\text{pseudo-}R^2 = \frac{1 - (L_0/L_1)^{2/n}}{1 - L_0^{2/n}} \tag{5.53}$$

$$= \frac{L_1^{2/n} - L_0^{2/n}}{1 - L_0^{2/n}} \tag{5.54}$$

where $n$ is the sample size. The minimum value of this quantity is zero when the fit is as bad as it can be (when $L_1 = L_0$), and the maximum value is one when the fit is as good as it can be (when $L_1 = 1$). This definition of pseudo-$R^2$ was suggested by Cragg and Uhler (1970). Unfortunately there is no formal significance test that utilizes this measure.

Another definition suggested by McFadden (1974) is simply

$$\text{pseudo-}R^2 = 1 - (\log L_0/\log L_1)^{2/n} \tag{5.55}$$

There is no formal test that utilizes this measure either.

The definition of pseudo-$R^2$ used by the SAS program for logit regression is

$$\boxed{\text{pseudo-}R^2 = \frac{2\log L_1 - 2\log L_0 - 2k}{-2\log L_0}} \tag{5.56}$$

where $k$ denotes the number of coefficients to be estimated, not counting the intercept. Recall from the previous section that the quantity $-2\log L_0 - 2\log L_1$ is model chi-square. If model chi-square is less than $2k$ in (5.56), pseudo-$R^2$ is set to zero (Harrell, 1983). Pseudo-$R$, computed as the square root of pseudo-$R^2$, is sometimes presented instead of pseudo-$R^2$. Pseudo-$R$, computed for the model presented earlier in equation (5.21), is .22.

There are several difficulties with these measures of pseudo-$R^2$, some of which have already been mentioned. First, there are several different measures available, which can give rather different numerical results when applied to the same data set. Second, there is little basis for choosing one measure over the other. Third, statistical tests that utilize pseudo-$R^2$ are not available for any of the measures. For these reasons, many authors do not present values of pseudo-$R^2$ when reporting results from logit regression analyses. If the analyst does use one of the measures, he or she should report which one.

## 5.5. MCA ADAPTED TO LOGIT REGRESSION

When multiple classification analysis (MCA) is adapted to logit regression, both unadjusted and adjusted values of the response variable can be calculated, just as in ordinary MCA. The unadjusted values are based on logit regressions that incorporate one predictor variable at a time, and the adjusted values are based on the complete model including all predictor variables simultaneously. We shall consider only adjusted values in this chapter.

Although pseudo-$R$ is the analogue of ordinary $R$, there is no logit regression analogue to partial $R$. Therefore, measures of partial association are omitted from the MCA tables based on logit regression.

Statistical computing packages such as SPSS, SAS, BMDP, and LIMDEP do not include MCA programs for logit regression. The analyst must construct the MCA tables from the underlying logit regressions, as described below.

### 5.5.1. An Illustrative Example

To illustrate how MCA can be adapted to logit regression, we consider the contraceptive use effect of age, age-squared, education (low, medium, and high), residence (urban, rural), ethnicity (Indians, Fijians), and residence × ethnicity. Whereas we previously considered fecund, nonpregnant, currently married women aged 35–44, we now broaden the age range to 15–49 while at the same time introducing a control for age and age-squared. An age-squared term is included because contraceptive use tends to increase with age up to about age 35 or 40, after which it tends to level off or even decline. This kind of simple curvature is readily captured by a quadratic term, as discussed earlier in Chapter 2.

Our variables are

$P$: estimated probability of contraceptive use

$A$: age

$M$: 1 if medium education, 0 otherwise
$H$: 1 if high education, 0 otherwise
$U$: 1 if urban, 0 otherwise
$I$: 1 if Indian, 0 otherwise

In log odds form, the model is

$$\log \Omega = a + bA + cA^2 + dM + fH + gU + hI + jUI \quad (5.57)$$

In odds form, the model is

$$\Omega = e^{a+bA+cA^2+dM+fH+gU+hI+jUI} \quad (5.58)$$

In probability form, the model is

$$P = \frac{1}{1 + e^{-(a+bA+cA^2+dM+fH+gU+hI+jUI)}} \quad (5.59)$$

When the model is fitted to the Fiji data, it is found, as shown in Table 5.1, that all coefficients except $j$ (the coefficient of the $UI$

**TABLE 5.1. Coefficients and Effects (Odds Ratios) for the Estimated Model, $\log \Omega = a + bA + cA^2 + dM + fH + gU + hI + jUI$: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | Coefficient (S.E.) | Effect (Odds Ratio) |
|---|---|---|
| Intercept | $-5.594$*(.610) | |
| Age | | $e^{.255+2(-.003)A} = 1.290\, e^{-.006A}$ |
| $A$ | .255*(.038) | — |
| $A^2$ | $-.003$*(.001) | — |
| Education | | |
| $H$ | .331*(.102) | $e^{.331} = 1.392$ |
| $M$ | .254*(.096) | $e^{.254} = 1.289$ |
| Residence × ethnicity | | |
| $U$ | .294*(.127) | $e^{.294-.046I} = 1.342\, e^{-.046I}$ |
| $I$ | 1.160*(.099) | $e^{1.160-.046U} = 3.190\, e^{-.046U}$ |
| $UI$ | $-.046$ (.158) | — |

[a]The response variable is the log odds of contraceptive use. An asterisk after a coefficient indicates that the coefficient differs from zero with a two-sided $p < .05$. Numbers in parentheses following coefficients are standard errors of coefficients. The effect of age depends on the level of $A$, the effect of residence depends on the level of $I$, and the effect of ethnicity depends on the level of $U$; if $A$, $U$, and $I$ are set to their mean values of 31.03, .353, and .609, respectively, these effects are 1.071 for age, 1.305 for residence ($U$), and 3.139 for ethnicity ($I$).

interaction) are significant at the 5 percent level or higher. Although the *UI* term could be dropped from the model (the model would then have to be reestimated without the *UI* term), we retain it for purposes of illustration. Table 5.1 also shows effects, as measured by odds ratios.

### 5.5.2. Table Set-up and Numerical Results

Table 5.2 shows the MCA table set-up and numerical results for adjusted values of log $\Omega$. The model is given by equation (5.57), which is repeated for convenience in the footnote to the table. Because (5.57) is in the form of a multiple regression equation, the table set-up is done in exactly the same way as in ordinary multiple regression, as described in Chapter 3. Although unadjusted values are not shown here, they would also be set up as in Chapter 3, had we included them.

Because age is a continuous variable, one must enter selected values of age, which play the role of categories. We have chosen 15, 25, 35, and 45. All the other variables are categorical to begin with and require no selection of values. In each panel, other variables are controlled by setting these variables at their mean values in the entire sample, and these mean values are given in the footnote to the table. When the values of the coefficients and mean values are substituted into the formulae in the first column of the table, one obtains the numerical values given in the second column of the table. Using the information given in the footnote to the table, the reader can replicate the numerical values in the second column of the table as a check on understanding.

Whereas Table 5.2 gives adjusted values of log $\Omega$, Table 5.3 gives corresponding adjusted values of $\Omega$ and $P$. The adjusted values of $\Omega$ are obtained by taking the adjusted values of log $\Omega$ in Table 5.2 as a power of $e$, that is, $\Omega = e^{\log \Omega}$. Adjusted values of $P$ are then calculated as

$$P = \frac{\Omega}{1 + \Omega} \tag{5.60}$$

This formula is derived by solving the equation $\Omega \equiv P/(1 - P)$ for $P$.

The numerical results in Tables 5.2 and 5.3 show that there is indeed a tendency for contraceptive use to level off at the older reproductive ages, which justifies the inclusion of the age-squared term in the model. As mentioned earlier, the interaction between

**TABLE 5.2. MCA Table of Adjusted Values of the Log Odds of Contraceptive Use for the Estimated Model, $\log \Omega = a + bA + cA^2 + dM + fH + gU + hI + jUI$: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a–c]

| Predictor Variable | $n$ | Adjusted Value of $\log \Omega$: Formula | Numerical Value |
|---|---|---|---|
| Age | | | |
| 15 | — | $a + 15b + (15)^2c + d\bar{M} + f\bar{H} + g\bar{U} + h\bar{I} + j\overline{UI}$ | −1.491 |
| 25 | — | $a + 25b + (25)^2c + d\bar{M} + f\bar{H} + g\bar{U} + h\bar{I} + j\overline{UI}$ | −.204 |
| 35 | — | $a + 35b + (35)^2c + d\bar{M} + f\bar{H} + g\bar{U} + h\bar{I} + j\overline{UI}$ | .451 |
| 45 | — | $a + 45b + (45)^2c + d\bar{M} + f\bar{H} + g\bar{U} + h\bar{I} + j\overline{UI}$ | .473 |
| Education | | | |
| Low | 1266 | $a + b\bar{A} + c\overline{A^2} + g\bar{U} + h\bar{I} + j\overline{UI}$ | −.106 |
| Medium | 1044 | $a + d + b\bar{A} + c\overline{A^2} + g\bar{U} + h\bar{I} + j\overline{UI}$ | .148 |
| High | 1169 | $a + f + b\bar{A} + c\overline{A^2} + g\bar{U} + h\bar{I} + j\overline{UI}$ | .226 |
| Residence and ethnicity | | | |
| Urban Indian | 854 | $a + g + h + j + b\bar{A} + c\overline{A^2} + d\bar{M} + f\bar{H}$ | .691 |
| Urban Fijian | 375 | $a + g + b\bar{A} + c\overline{A^2} + d\bar{M} + f\bar{H}$ | −.423 |
| Rural Indian | 1265 | $a + h + b\bar{A} + c\overline{A^2} + d\bar{M} + f\bar{H}$ | .443 |
| Rural Fijian | 985 | $a + b\bar{A} + c\overline{A^2} + d\bar{M} + f\bar{H}$ | −.717 |
| Total | 3479 | | |

[a]The estimated model is

$$\log \Omega = -5.594 + .255A - .003A^2 + .254M + .331H + .294U + 1.160I - .046UI$$

Mean values of the predictor variables are $\bar{A} = 31.03$, $\overline{A^2} = 1021.66$, $\bar{M} = .300$, $\bar{H} = .336$, $\bar{U} = .353$, $\bar{I} = .609$, and $\overline{UI} = .245$. Note that $A^2$ and $UI$ are treated as separate variables, so that means are calculated as the mean of $A^2$ (instead of the square of the mean of $A$) and the mean of $UI$ (instead of the mean of $U$ times the mean of $I$). $-2\log L_0 = 4817.20$ for the intercept model and $-2\log L_1 = 4462.12$ for the test model, given above. The test model differs significantly from the intercept model, with $p < .001$ from the likelihood ratio test. Pseudo-$R$ for the test model is .27.

[b]The values of $\log \Omega$ in the table were actually calculated from values of coefficients and means accurate to seven significant figures. It turns out that, in this example, the use of figures rounded to only three decimal places introduces substantial errors in the estimates of $\log \Omega$, $\Omega$, and $P$. The main source of these errors is the use of $-.003$ instead of $-.003162000$ for the coefficient of $A^2$.

[c]To be on the safe side, intermediate calculations should always employ data values with at least two more significant figures than the number of significant figures desired in the final estimate. Note that significant figures are not the same as decimal places. The data value .002 has three decimal places but only one significant figure, because leading zeros are not counted as significant figures. The data value 1.003 has three decimal places but four significant figures. The number 1.0030 has five significant figures, because the last zero indicates that the data value is accurate to four decimal places.

**TABLE 5.3. MCA Table of the Adjusted Effects of Age, Education, Residence, and Ethnicity on the Odds of Contraceptive Use and the Probability of Contraceptive Use for the Estimated Model, $\log \Omega = a + bA + cA^2 + dM + fM + gU + hI + jUI$: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | | Adjusted Value | |
|---|---|---|---|
| | $n$ | $\Omega$ | $P$ |
| Age | | | |
| 15 | — | .225 | .18 |
| 25 | — | .815 | .45 |
| 35 | — | 1.569 | .61 |
| 45 | — | 1.605 | .62 |
| Education | | | |
| Low | 1266 | .900 | .47 |
| Medium | 1044 | 1.160 | .54 |
| High | 1169 | 1.253 | .56 |
| Residence and ethnicity | | | |
| Urban Indian | 854 | 1.995 | .67 |
| Urban Fijian | 375 | .655 | .40 |
| Rural Indian | 1265 | 1.557 | .61 |
| Rural Fijian | 985 | .488 | .33 |
| Total | 3479 | | |

[a] Values of $\Omega$ are calculated by taking the values of $\log \Omega$ in Table 5.2 as a power of $e$. Values of $P$ are then calculated as $\Omega/(1 + \Omega)$. Among the 3479 women, 1810 are users and 1669 are nonusers. Values of $\Omega$ and $P$ were calculated from more accurate values of $\log \Omega$ than shown in Table 5.2; see footnote in Table 5.2.

residence and ethnicity is not statistically significant, although we have left the interaction in the model. Urban women have somewhat higher rates of contraceptive use than do rural women, controlling for the other variables in the model. Indian women have considerably higher rates of contraceptive use than do Fijian women, controlling for the other variables in the model.

Although the probabilities in the last column of Table 5.3 are presented in decimal form, they are frequently presented in the literature in percentage form. In percentage form, the value of $P$ for, say, urban Indians would be presented as 67 percent instead of .67.

Normally, in a published research paper in an academic journal, one would not present formulae for adjusted $\log \Omega$, as shown in Table 5.2. Instead, one would omit $\log \Omega$ altogether from the tables and combine Tables 5.2 and 5.3 into a single table, with columns for numerical values of $n$, adjusted $\Omega$, and adjusted $P$. Because adjusted $\Omega$ can be calculated from adjusted $P$ as $\Omega = P(1 - P)$, even adjusted $\Omega$ would usually be omitted.

## 5.6. FITTING THE LOGIT REGRESSION MODEL

To fit the logit regression model, we use the *method of maximum likelihood*.

### 5.6.1. The Maximum Likelihood Method

To illustrate this method, let us consider a very simple model:

$$\text{logit } P = a + bX \tag{5.61}$$

which can be written alternatively as

$$P = \frac{1}{1 + e^{-(a+bX)}} \tag{5.62}$$

We assume that the mathematical form of (5.61) and (5.62) is correct, but we don't yet know the values of $a$ and $b$, which are treated as unknowns.

The first step is to formulate a *likelihood function*, $L$. As mentioned earlier, $L$ is the probability of observing our particular sample data under the assumption that the model is true. That is, we assume that the mathematical form of the model, as given by (5.61) or (5.62), is correct, but we don't yet know the values of $a$ and $b$. *We choose a and b so that L is maximized. In other words, we choose values of the unknown parameters that maximize the likelihood of the observed data and call these the best-fitting parameters.* The method can be thought of as considering all possible combinations of $a$ and $b$, calculating $L$ for each combination, and picking the combination that yields the largest value of $L$. In practice we use a mathematical shortcut, which is described below.

### 5.6.2. Derivation of the Likelihood Function

Although we shall not usually be concerned about the mathematical details of maximum likelihood estimation, it is nevertheless useful to examine a simple example to provide a better feel for how the method works.

Suppose that our original response variable is $Y$, which for individuals is either 0 or 1. Our model is

$$P = \frac{1}{1 + e^{-(a+bX)}} \qquad \text{(5.62 repeated)}$$

Denote the first individual by subscript 1. For this individual $(X, Y) = (X_1, Y_1)$. We then have that

$$P_1 = \frac{1}{1 + e^{-(a+bX_1)}} \qquad (5.63)$$

Similarly, for individual 2,

$$P_2 = \frac{1}{1 + e^{-(a+bX_2)}} \qquad (5.64)$$

And so on for the remaining individuals in the sample.

Let us consider individual 1 in more detail. We have from (5.63) that

$$\Pr(Y_1 = 1) = P_1 \qquad (5.65)$$

where Pr denotes "probability that." Therefore,

$$\Pr(Y_1 = 0) = 1 - P_1 \qquad (5.66)$$

We can combine (5.65) and (5.66) into one formula:

$$\Pr(Y_1) = P_1^{Y_1}(1 - P_1)^{(1-Y_1)} \qquad (5.67)$$

Let us check that this equation works. If $Y_1 = 1$, then

$$\Pr(Y_1 = 1) = P_1^1(1 - P_1)^{(1-1)} = P_1 \qquad \textit{Checks} \qquad (5.68)$$

If $Y_1 = 0$, then

$$\Pr(Y_1 = 0) = P_1^0(1 - P_1)^{(1-0)} = 1 - P_1 \qquad \textit{Checks} \qquad (5.69)$$

Similarly

$$\Pr(Y_2) = P_2^{Y_2}(1 - P_2)^{(1-Y_2)} \qquad (5.70)$$

And so on up to $\Pr(Y_n)$, where $n$ denotes the number of sample cases.

If we assume simple random sampling, these $n$ probabilities are independent. Then the probability, or likelihood, of observing our particular sample data is

$$\begin{aligned}
\Pr(Y_1, Y_2, \ldots, Y_n) &= \Pr(Y_1)\Pr(Y_2)\ldots\Pr(Y_n) \\
&= \prod_{i=1}^{n} \Pr(Y_i) \\
&= \prod \left[ P_i^{Y_i}(1 - P_i)^{(1-Y_i)} \right] \\
&= \prod \left[ \frac{1}{1 + e^{-(a+bX_i)}} \right]^{Y_i} \left[ 1 - \frac{1}{1 + e^{-(a+bX_i)}} \right]^{1-Y_i} \\
&= L(a, b) \qquad (5.71)
\end{aligned}$$

where $\prod$ is the product symbol (analogous to the summation symbol $\Sigma$), and where $L(a, b)$ indicates that the likelihood function, $L$, is a function of the unknown parameters, $a$ and $b$. Note that the $X_i$ and $Y_i$ are observed data and therefore constants in the equation, not unknown parameters.

It is also useful to derive a formula for $\log L$:

$$\log L(a, b) = \sum \left\{ Y_i \log\left[ \frac{1}{1 + e^{-(a+bX_i)}} \right] + (1 - Y_i)\log\left[ 1 - \frac{1}{1 + e^{-(a+bX_i)}} \right] \right\} \qquad (5.72)$$

In getting from (5.71) to (5.72), we make use of the rules $\log xy = \log x + \log y$ and $\log x^y = y \log x$.

We wish to find the values of $a$ and $b$ that maximize $L(a, b)$. Because $\log L$ is a monotonically increasing function of $L$, maximizing $L$ is equivalent to maximizing $\log L$. In other words, the same values of $a$ and $b$ that maximize $L$ also maximize $\log L$. Mathematically, it is easier to maximize $\log L$ than to maximize $L$ directly. We maximize $\log L$ by taking the partial derivative of $\log L$ first with respect to $a$ and second with respect to $b$, and then equating each derivative separately to zero, yielding two equations in two unknowns, $a$ and $b$. These equations are then solved for $a$ and $b$ by numerical methods that are beyond the scope of this book. Derivation of the standard errors of $a$ and $b$ is also beyond the scope of this book.

We have sketched the essentials of this application of the maximum likelihood estimation (MLE) method in order to give the reader a feel for how the method works. MLE is also used to fit the multinomial logit model and the hazard model, which are considered in later chapters. Henceforth, we shall simply pronounce the magic words "maximum likelihood" and accept on faith that the model has been fitted. We shall then focus on model specification and interpretation, as in earlier chapters.

## 5.7. SOME LIMITATIONS OF THE LOGIT REGRESSION MODEL

In logit regression, the lower and upper asymptotes of the logistic function for $P$ are 0 and 1. In some applications, this may not be realistic. Suppose, for example, that we are looking at the effect of education on contraceptive use in the United States. In the United States, even persons with no education have a probability of using contraception that is well above zero, so that a lower asymptote of zero may be unrealistic. This is not serious a problem as it might seem, however, because when we fit the logit regression model, we do not use the entire curve. In effect, we use only a piece of the curve corresponding to the observed range of the predictor variable.

## 5.8. FURTHER READING

A simple discussion of logit regression may be found in Wonnacott and Wonnacott (1979), *Econometrics*. A somewhat more advanced treatment is found in Hanushek and Jackson (1977), *Statistical Methods for Social Scientists* and in Fox (1984), *Linear Statistical Models and Related Methods*. For more advanced treatments, see Walker and Duncan (1967), Cox (1970), Amemiya (1981), and Maddala (1983).

# 6

# MULTINOMIAL LOGIT REGRESSION

The *multinomial logit model* (also called the *polytomous logit model*) is a generalization of the binary logit model considered in Chapter 5. In this context, "binary" means that the response variable has two categories, and "multinomial" means that the response variable has three or more categories. As in binary logit regression, the predictors in multinomial logit regression may be quantitative, categorical, or a mixture of the two.

## 6.1. FROM LOGIT TO MULTINOMIAL LOGIT

### 6.1.1. The Basic Form of the Multinomial Logit Model

The explication of the multinomial logit model is facilitated by a simple example. Suppose that the response variable is contraceptive method choice:

$P_1$: estimated probability of using sterilization
$P_2$: estimated probability of using some other method
$P_3$: estimated probability of using no method

The categories of the response variable are mutually exclusive and exhaustive: A sample member must fall in one and only one of the categories.

Suppose that the reference category is "no method." As in previous chapters, the choice of reference category is arbitrary, insofar as this choice has no effect on the final estimated probabilities of using each method.

Suppose also that the predictor variables are education (low, medium, or high) and ethnicity (Indian or Fijian):

$M$: 1 if medium education, 0 otherwise
$H$: 1 if high education, 0 otherwise
$I$: 1 if Indian, 0 otherwise

Our theory is that education and ethnicity influence contraceptive method choice.

The multinomial logit model then consists of two equations plus a constraint:

$$\log\frac{P_1}{P_3} = a_1 + b_1M + c_1H + d_1I \qquad (6.1a)$$

$$\log\frac{P_2}{P_3} = a_2 + b_2M + c_2H + d_2I \qquad (6.1b)$$

$$P_1 + P_2 + P_3 = 1 \qquad (6.1c)$$

Logarithms are natural logarithms (base $e$). In general, the number of model equations (including the constraint) equals the number of categories of the response variable.

Strictly speaking, the quantities $P_1/P_3$ and $P_2/P_3$ in (6.1) are not odds, because numerator and denominator do not necessarily sum to one. We may think of $P_1/P_3$ and $P_2/P_3$ as "improper" odds. For convenience, however, and in accordance with common usage, we shall refer to them simply as odds. Each of these odds has for its denominator the probability of the reference category of the response variable.

The model in (6.1) can be fitted by the method of maximum likelihood. We assume that the mathematical form of the model is correct, and we choose the values of $a_1$, $b_1$, $c_1$, $d_1$, $a_2$, $b_2$, $c_2$, and $d_2$ to maximize the likelihood function. We shall not worry about the mathematical details of how this is done. Instead, we shall assume that the model has been fitted and proceed to examine how to use and interpret it.

### 6.1.2. Interpretation of Coefficients

In multinomial logit regression, the interpretation of coefficients is not as straightforward as in binary logit regression.

Suppose, for example, that $d_1$ is positive in (6.1a). Then a one-unit increase in $I$ (from 0 to 1) causes $\log(P_1/P_3)$ to increase by $d_1$ units. When $\log(P_1/P_3)$ increases, the odds $P_1/P_3$ also increase, because $\log(P_1/P_3)$ is a monotonic-increasing function of $P_1/P_3$.

However, we cannot reason that $P_1$ itself increases. $P_1$ could actually decrease. This could happen if $P_3$ also decreases and if the proportionate decrease in $P_3$ exceeds the proportionate decrease in $P_1$. In sum, $P_1/P_3$ can increase while both $P_1$ and $P_3$ decrease.

*Therefore, a positive value of $d_1$ does not necessarily mean that a one-unit increase in $I$ acts to increase $P_1$. The opposite may be true.*

This could not happen in binary logit regression, because the numerator and denominator of $P/(1 - P)$ cannot move in the same direction. If $P$ increases, $1 - P$ must decrease by the same amount. Therefore, if $P/(1 - P)$ increases, $P$ must also increase.

The above discussion illustrates that, in multinomial logit regression, the effects of the predictor variables on $\log(P_1/P_3)$ and $P_1/P_3$ can be misleading, because the effects on $P_1$ can be in the opposite direction. The same point applies to the effects of the predictor variables on $\log(P_2/P_3)$, $P_2/P_3$, and $P_2$. Therefore, in presenting results of multinomial logit analysis, we deemphasize the odds and log odds and focus instead on the effects of the predictor variables directly on $P_1$, $P_2$, and $P_3$.

### 6.1.3. Presentation of Results in Multiple Classification Analysis (MCA) Format

As in binary logit regression, the most convenient way to present the effects of the predictor variables on $P_1$, $P_2$, and $P_3$ is in the form of an MCA table, which is constructed in the following way:

The first step is to take each side of (6.1a) and (6.1b) as a power of $e$ and then multiply through by $P_3$, yielding

$$P_1 = P_3 e^{a_1 + b_1 M + c_1 H + d_1 I} \tag{6.2a}$$

$$P_2 = P_3 e^{a_2 + b_2 M + c_2 H + d_2 I} \tag{6.2b}$$

We also have the identity

$$P_3 = P_3 \tag{6.2c}$$

If we now add (6.2a), (6.2b), and (6.2c) and recall that $P_1 + P_2 + P_3 = 1$, we get

$$1 = P_3 \sum_{j=1}^{2} e^{a_j + b_j M + c_j H + d_j I} + P_3 \tag{6.3}$$

Solving (6.3) for $P_3$, we obtain

$$P_3 = \frac{1}{1 + \sum e^{a_j + b_j M + c_j H + d_j I}} \tag{6.4}$$

where the notation for the summation is now abbreviated to omit the limits of summation.

Substituting (6.4) back into (6.2a) and (6.2b) and repeating (6.4), we obtain

$$P_1 = \frac{e^{a_1 + b_1 M + c_1 H + d_1 I}}{1 + \sum e^{a_j + b_j M + c_j H + d_j I}} \tag{6.5a}$$

$$P_2 = \frac{e^{a_2 + b_2 M + c_2 H + d_2 I}}{1 + \sum e^{a_j + b_j M + c_j H + d_j I}} \tag{6.5b}$$

$$P_3 = \frac{1}{1 + \sum e^{a_j + b_j M + c_j H + d_j I}} \tag{6.5c}$$

where the summations range from $j = 1$ to $j = 2$. Equations (6.5), which are an alternative statement of the model in equations (6.1), are calculation formulae for $P_1$, $P_2$, and $P_3$. A shortcut for calculating $P_3$ is to calculate $P_1$ and $P_2$ from (6.5a) and (6.5b) and then to obtain $P_3$ as $1 - (P_1 + P_2)$.

The MCA table is constructed by substituting appropriate combinations of ones, zeros, and mean values in equations (6.5), as shown in Table 6.1. For example, the formulae for $P_1$, $P_2$, and $P_3$ for those with high education in Table 6.1 are obtained by substituting $M = 0$, $H = 1$, and $I = \bar{I}$ in (6.5a), (6.5b), and (6.5c). The formulae for $P_1$, $P_2$, and $P_3$ for Fijians are obtained by substituting $I = 0$, $M = \bar{M}$, and $H = \bar{H}$ in (6.5a), (6.5b), and (6.5c).

The calculation of the MCA table for multinomial logit regression can be very tedious on a desk calculator. The calculation is most easily accomplished by transferring the multinomial logit computer output (fitted coefficients along with the mean values of the predictor variables) into a spreadsheet program like LOTUS.

**TABLE 6.1. Set-up for MCA Table of Adjusted Values of $P_j$ for the Model $\log P_j / P_3 = a_j + b_j M + c_j H + d_j I$, $j = 1, 2$: Fecund, Nonpregnant, Currently Married Women Aged 35–44 in the 1974 Fiji Fertility Survey[a]**

| Predictor Variable | $n$ | $P_1$ (Sterilization) | $P_2$ (Other Methods) | $P_3$ (No Method) |
|---|---|---|---|---|
| Education | | | | |
| Low | $n_L$ | $\frac{e^{a_1+d_1\bar{I}}}{1+\sum e^{a_j+d_j\bar{I}}}$ | $\frac{e^{a_2+d_2\bar{I}}}{1+\sum e^{a_j+d_j\bar{I}}}$ | $\frac{1}{1+\sum e^{a_j+d_j\bar{I}}}$ |
| Medium | $n_M$ | $\frac{e^{a_1+b_1+d_1\bar{I}}}{1+\sum e^{a_j+b_j+d_j\bar{I}}}$ | $\frac{e^{a_2+b_2+d_2\bar{I}}}{1+\sum e^{a_j+b_j+d_j\bar{I}}}$ | $\frac{1}{1+\sum e^{a_j+b_j+d_j\bar{I}}}$ |
| High | $n_H$ | $\frac{e^{a_1+c_1+d_1\bar{I}}}{1+\sum e^{a_j+c_j+d_j\bar{I}}}$ | $\frac{e^{a_2+c_2+d_2\bar{I}}}{1+\sum e^{a_j+c_j+d_j\bar{I}}}$ | $\frac{1}{1+\sum e^{a_j+c_j+d_j\bar{I}}}$ |
| Ethnicity | | | | |
| Indian | $n_I$ | $\frac{e^{a_1+d_1+b_1\bar{M}+c_1\bar{H}}}{1+\sum e^{a_j+d_j+b_j\bar{M}+c_j\bar{H}}}$ | $\frac{e^{a_2+d_2+b_2\bar{M}+c_2\bar{H}}}{1+\sum e^{a_j+d_j+b_j\bar{M}+c_j\bar{H}}}$ | $\frac{1}{1+\sum e^{a_j+d_j+b_j\bar{M}+c_j\bar{H}}}$ |
| Fijian | $n_F$ | $\frac{e^{a_1+b_1\bar{M}+c_1\bar{H}}}{1+\sum e^{a_j+b_j\bar{M}+c_j\bar{H}}}$ | $\frac{e^{a_2+b_2\bar{M}+c_2\bar{H}}}{1+\sum e^{a_j+b_j\bar{M}+c_j\bar{H}}}$ | $\frac{1}{1+\sum e^{a_j+b_j\bar{M}+c_j\bar{H}}}$ |
| $n$ | $n_T$ | $n_S$ | $n_O$ | $n_N$ |

[a]Summations in the denominators range from $j = 1$ to $j = 2$. Calculation formulae are derived from equations (6.5) in the text. In the last line of the table, $n_T$ denotes total sample size, $n_S$ denotes the sample number of sterilized women, $n_O$ denotes the sample number using some other method, and $n_N$ denotes the sample number using no method.

The results of fitting this model to data from the 1974 Fiji Fertility Survey are shown in Tables 6.2 and 6.3. In Table 6.3 the probabilities are multiplied by 100 and thereby reexpressed as percentages.

Table 6.2 shows the estimated coefficients and their standard errors, which are printed out by the multinomial logit program. The coefficients for equation (6.1a) are given in the column labeled $\log(P_1/P_3)$, and the coefficients for equation (6.1b) are given in the column labeled $\log(P_2/P_3)$. Table 6.3 transforms the results into an MCA format by applying the formulae in Table 6.1 and utilizing the values of the coefficients in Table 6.2 and the mean values of the variables given in the footnote to Table 6.3. Using the formulae in Table 6.1, the coefficients in Table 6.2, and the mean values of the predictor variables given in the footnote to Table 6.3, the reader should recalculate the probabilities in Table 6.3 as a check on understanding.

**TABLE 6.2. Multinomial Logit Regression Coefficients for the Model $\log P_j / P_3 = a_j + b_j M + c_j H + d_j I$, $j = 1, 2$: Fecund, Nonpregnant, Currently Married Women Aged 35–44 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | $\log(P_1/P_3)$ | $\log(P_2/P_3)$ |
|---|---|---|
| Intercept | −1.299* (.197) | −1.153* (.197) |
| Education | | |
| Medium ($M$) | .373 (.203) | .489* (.216) |
| High ($H$) | .169 (.235) | .709* (.234) |
| Ethnicity | | |
| Indian ($I$) | 1.636* (.192) | .732* (.193) |

[a]The underlying model is given in equations (6.1). The three methods are (1) sterilization, (2) other method, and (3) no method. $P_1/P_3$ compares sterilization with no method, and $P_2/P_3$ compares other method with no method. Numbers in the $\log(P_1/P_3)$ column are $a_1$, $b_1$, $c_1$, and $d_1$, and numbers in the $\log(P_2/P_3)$ column are $a_2$, $b_2$, $c_2$, and $d_2$ [see equations (6.1)]. An asterisk after a coefficient indicates that the coefficient differs from zero with a two-tailed $p < .05$. Numbers in parentheses following coefficients are standard errors. The log likelihood of the test model is $\log L_1 = -982.25$. The log likelihood of the intercept model is $\log L_0 = -1031.20$. The likelihood ratio test of the difference between the test model and the intercept model (see Section 5.3.2 in Chapter 5) yields $\chi^2 = 97.9$ with d.f. = 6, implying that $p < .001$ [see the $\chi^2$ table in Appendix B (Table B.4)]. Pseudo-$R$ = .20.

**TABLE 6.3. MCA Table of Adjusted Values of $P_j$ (in Percent) for the Model $\log P_j / P_3 = a_j + b_j M + c_j H + d_j I$, $j = 1, 2$: Fecund, Nonpregnant, Currently Married Women Aged 35–44 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | $n$ | $P_1$ (Sterilization) | $P_2$ (Other Method) | $P_3$ (No Method) |
|---|---|---|---|---|
| Education | | | | |
| Low | 386 | 33 | 22 | 45 |
| Medium | 289 | 37 | 28 | 35 |
| High | 179 | 30 | 35 | 35 |
| Ethnicity | | | | |
| Indian | 575 | 46 | 24 | 29 |
| Fijian | 379 | 18 | 23 | 59 |
| $n$ | 954 | 335 | 239 | 380 |

[a]Values in this table are obtained using the calculation formulae in Table 6.1, the estimates of coefficients in Table 6.2, and the following mean values of the predictor variables: $\bar{M} = .303$, $\bar{H} = .188$, and $\bar{I} = .603$.

### 6.1.4. Statistical Inference

Tests of coefficients, tests of the difference between two coefficients, and tests of the difference between two models are done in the same way as in binary logit regression. (See Section 5.3 of Chapter 5.) Some of these tests are illustrated in Table 6.2.

Because the sign of a multinomial logit regression coefficient may not reflect the direction of effect of the predictor variable on either of the probabilities on the left-hand side of the equation, tests of coefficients are usually two-tailed tests.

When testing the difference between two models, one of which is the intercept model, the form of the intercept model, referring back to our earlier example, is

$$\log\frac{P_1}{P_3} = a_1 \qquad (6.6a)$$

$$\log\frac{P_2}{P_3} = a_2 \qquad (6.6b)$$

$$P_1 + P_2 + P_3 = 1 \qquad (6.6c)$$

### 6.1.5. Goodness of Fit

Pseudo-$R^2$ is calculated in the same way as in binary logit regression. [See equation (5.56) in Section 5.4 of Chapter 5.] In the multinomial logit example in Tables 6.1–6.3, $k = 6$ and the value of pseudo-$R$ is .20 (see footnote in Table 6.2). The reader can check this result by substituting the log likelihood statistics into equation (5.56).

### 6.1.6. Changing the Reference Category of the Response Variable

Suppose we ran the model in (6.1) with "other method" instead of "no method" as the reference category. The model would be

$$\log\frac{P_1}{P_2} = e_1 + f_1M + g_1H + h_1I \qquad (6.7a)$$

$$\log\frac{P_3}{P_2} = e_2 + f_2M + g_2H + h_2I \qquad (6.7b)$$

$$P_1 + P_2 + P_3 = 1 \qquad (6.7c)$$

As already mentioned, the estimated values of $P_1$, $P_2$, and $P_3$ for each specified contribution of values of the predictor variables would come out the same as before. However, the coefficients would be different.

It is not necessary to rerun the model with "other method" as the new reference category to estimate and test the coefficients in equations (6.7). Instead, we can derive (6.7) from (6.1). Subtracting equation (6.1b) from (6.1a) and making use of the rule that $\log A - \log B = \log(A/B)$, we obtain

$$\log\frac{P_1}{P_2} = (a_1 - a_2) + (b_1 - b_2)M + (c_1 - c_2)H + (d_1 - d_2)I \tag{6.8a}$$

Comparing (6.7a) and (6.8a), we see that $e_1 = a_1 - a_2$, $f_1 = b_1 - b_2$, $g_1 = c_1 - c_2$, and $h_1 = d_1 - d_2$.

Multiplying both sides of (6.1b) by $-1$ and making use of the rule that $-\log A = \log(1/A)$, we obtain

$$\log\frac{P_3}{P_2} = -a_2 - b_2M - c_2H - d_2I \tag{6.8b}$$

Comparing (6.7b) and (6.8b), we see that $e_2 = -a_2$, $f_2 = -b_2$, $g_2 = -c_2$, and $h_2 = -d_2$.

In sum, when we change the reference category of the response variable, we can obtain the new coefficients from the original coefficients without rerunning the model.

We can also test the significance of the new coefficients without rerunning the model. In (6.7a) and (6.8a), for example, we wish to test whether the coefficient of $M$, $f_1 = b_1 - b_2$, differs significantly from zero. In multinomial logit regression as in binary logit regression and ordinary multiple regression, one can optionally print out a regression coefficient covariance matrix, which enables one to compute the standard error of $b_1 - b_2$ as

$$\sqrt{\mathrm{Var}(b_1 - b_2)} = \sqrt{\mathrm{Var}(b_1) + \mathrm{Var}(b_2) - 2\,\mathrm{Cov}(b_1, b_2)}\,.$$

It is then a simple matter to test whether $b_1 - b_2$ differs significantly from zero. The test procedure for the difference between two coefficients is the same as that described earlier in Section 2.9.2 of Chapter 2.

## 6.2. MULTINOMIAL LOGIT MODELS WITH INTERACTIONS AND NONLINEARITIES

As a more complicated example including interactions and nonlinearities, let us consider the effects of age, age-squared, education (low, medium, high), residence (urban, rural), ethnicity (Indian, Fijian), and residence × ethnicity on contraceptive method choice among fecund, nonpregnant, currently married women aged 15–49. This example is identical to that given in Section 5.5.1 of Chapter 5, except that the response variable is now contraceptive method choice among three methods (sterilization, other method, no method) instead of two methods (use, non-use).

The predictor variables are

$A$: age
$M$: 1 if medium education, 0 otherwise
$H$: 1 if high education, 0 otherwise
$U$: 1 if urban, 0 otherwise
$I$: 1 if Indian, 0 otherwise

In log odds form, the model is

$$\log\frac{P_1}{P_3} = a_1 + b_1A + c_1A^2 + d_1M + f_1H + g_1U + h_1I + i_1UI \quad (6.9a)$$

$$\log\frac{P_2}{P_3} = a_2 + b_2A + c_2A^2 + d_2M + f_2H + g_2U + h_2I + i_2UI \quad (6.9b)$$

$$P_1 + P_2 + P_3 = 1 \quad (6.9c)$$

In probability form, the model is

$$P_1 = \frac{e^{a_1+b_1A+c_1A^2+d_1M+f_1H+g_1U+h_1I+i_1UI}}{1+\sum e^{a_j+b_jA+c_jA^2+d_jM+f_jH+g_jU+h_jI+i_jUI}} \quad (6.10a)$$

$$P_2 = \frac{e^{a_2+b_2A+c_2A^2+d_2M+f_2H+g_2U+h_2I+i_2UI}}{1+\sum e^{a_j+b_jA+c_jA^2+d_jM+f_jH+g_jU+h_jI+i_jUI}} \quad (6.10b)$$

$$P_3 = \frac{1}{1+\sum e^{a_j+b_jA+c_jA^2+d_jM+f_jH+g_jU+h_jI+i_jUI}} \quad (6.10c)$$

where the summations range from $j = 1$ to $j = 2$.

**TABLE 6.4. Multinomial Logit Regression Coefficients for the Model $\log P_j / P_3 = a_j + b_j A + c_j A^2 + d_j M + f_j H + g_j U + h_j I + i_j UI,\ j = 1,2$: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | $\log(P_1/P_3)$ | $\log(P_2/P_3)$ |
|---|---|---|
| Intercept | −18.247* (1.311) | −4.304* (.661) |
| Age ($A$) | .856* ( .074) | .187* (.042) |
| Age-squared ($A^2$) | −.011* ( .001) | −.003* (.001) |
| Medium education ($M$) | .157 ( .127) | .372* (.109) |
| High education ($H$) | −.155 ( .146) | .552* (.112) |
| Urban ($U$) | .869* ( .193) | .050 (.144) |
| Indian ($I$) | 1.740* ( .150) | .899* (.110) |
| Urban × Indian ($UI$) | −.710* ( .228) | .255 (.178) |

[a]The underlying model is given in equations (6.9). The three methods are sterilization, other method, and no method. $P_1/P_3$ compares sterilization with no method, and $P_2/P_3$ compares other method with no method. An asterisk after a coefficient indicates that the coefficient differs from zero with a two-sided $p < .05$. Numbers in parentheses following coefficients are standard errors. The log likelihood of the test model is $\log L_1 = -3199.44$. The log likelihood of the intercept model is $\log L_0 = -3623.01$. The likelihood ratio test of the difference between the test model and the intercept model is 847.14 with d.f. = 14; consultation of the $\chi^2$ table in Appendix B (Table B.4) indicates $p < .001$. Pseudo-$R = .34$.

The MCA table is constructed by substituting appropriate combinations of ones, zeros, and mean values in equations (6.10). Because $A^2$ and $UI$ are treated as separate variables, means are calculated as the mean of $A^2$ (instead of the square of the mean of $A$) and the mean of $UI$ (instead of the mean of $U$ times the mean of $I$).

Results of fitting the model, as given in equations (6.9) and (6.10), are shown in Tables 6.4 and 6.5. As a check on understanding, the reader should derive the numbers in Table 6.5, starting from equations (6.10), the fitted coefficients in Table 6.4, and the mean values of the predictor variables in the footnote to Table 6.5.

## 6.3. A MORE GENERAL FORMULATION OF THE MULTINOMIAL LOGIT MODEL

We now consider the more general case where the response variable has $J$ mutually exclusive and exhaustive categories, denoted $j = 1, 2, \ldots, J$. The $J$th category is taken as the reference category for the response variable. Because the ordering of the categories is arbitrary,

**TABLE 6.5. MCA Table of Adjusted Values of $P_j$ (in Percent) for the Model $\log P_j / P_3 = a_j + b_j A + c_j A^2 + d_j M + f_j H + g_j U + h_j I + i_j UI$, $j = 1, 2$: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | $n$ | $P_1$ (Sterilization) | $P_2$ (Other Method) | $P_3$ (No Method) |
|---|---|---|---|---|
| Age | | | | |
| 15 | — | 0 | 23 | 77 |
| 25 | — | 6 | 37 | 57 |
| 35 | — | 31 | 30 | 38 |
| 45 | — | 35 | 23 | 42 |
| Education | | | | |
| Low | 1266 | 14 | 28 | 58 |
| Medium | 1044 | 14 | 36 | 50 |
| High | 1169 | 10 | 41 | 48 |
| Residence × ethnicity | | | | |
| Urban Indian | 854 | 18 | 44 | 37 |
| Urban Fijian | 375 | 11 | 24 | 65 |
| Rural Indian | 1265 | 18 | 38 | 44 |
| Rural Fijian | 985 | 5 | 25 | 70 |
| $n$ | 3479 | 715 | 1095 | 1669 |

[a]Values in this table are obtained using equations (6.10), the estimates of coefficients in Table 6.4, and the following means of the predictor variables: $\bar{A} = 31.03$, $\overline{A^2} = 1021.66$, $\bar{M} = .300$, $\bar{H} = .336$, $\bar{U} = .353$, $\bar{I} = .609$, and $\overline{UI} = .245$.

any category can be the $J$th category, so that the choice of the reference category is also arbitrary.

In the general case there are also $K$ predictor variables, denoted $X_1, X_2, \ldots, X_K$. The variables $X_i$ may denote not only variables like $A$, $U$, and $I$, but also variables like $A^2$ and $UI$.

The multinomial logit model is then specified in log odds form as

$$\log \frac{P_j}{P_J} = \sum_k b_{jk} X_k, \qquad j = 1, 2, \ldots, J-1 \qquad (6.11a)$$

$$\sum_j P_j = 1 \qquad (6.11b)$$

where the summation $\Sigma_k$ ranges from $k = 0$ to $k = K$; where, for convenience, $X_0$ is defined as $X_0 \equiv 1$; and where the summation $\Sigma_j$ ranges from $j = 1$ to $j = J$. We define $X_0 \equiv 1$ in order to be able to write the right side of (6.11a) more compactly as a summation of a single term, $b_{jk}X_k$, that includes the intercept as $b_{j0} = b_{j0}X_0$. Equation (6.11a) actually represents $J - 1$ equations. Therefore, equations (6.11a) and (6.11b) together represent $J$ equations, with $(J - 1)(K + 1)$ coefficients to be estimated.

By taking each side of (6.11a) as a power of $e$ and multiplying through by $P_J$, we can rewrite (6.11a) as

$$P_j = P_J e^{\Sigma_k b_{jk} X_k}, \qquad j = 1, 2, \ldots, J - 1 \tag{6.12a}$$

We also have

$$P_J = P_J \tag{6.12b}$$

We now sum the $J$ equations in (6.12a) and (6.12b) to obtain

$$1 = P_J \sum_j e^{\Sigma_k b_{jk} X_k} + P_J \tag{6.13}$$

where $X_0 = 1$, the summation over $j$ ranges from $j = 1$ to $j = J - 1$, and the summation over $k$ ranges from $k = 0$ to $k = K$. Solving for $P_J$, we get

$$P_J = \frac{1}{1 + \Sigma_j e^{\Sigma_k b_{jk} X_k}} \tag{6.14}$$

Substituting back into (6.12a) and (6.12b), we obtain the formulation of the model in probability form:

$$\boxed{P_j = \frac{e^{\Sigma_k b_{jk} X_k}}{1 + \Sigma_i e^{\Sigma_k b_{ik} X_k}}, \qquad j = 1, 2, \ldots, J} \tag{6.15}$$

where $X_0 = 1$, the summation $\Sigma_k$ ranges from $k = 0$ to $k = K$, the summation $\Sigma_i$ ranges from $i = 1$ to $i = J - 1$, and $b_{J0}, b_{J1}, \ldots, b_{JK}$ are all defined to be zero. This latter definition implies that $e^{\Sigma_k b_{Jk} X_k} = e^0 = 1$, so that (6.15) reduces to (6.14) when $j = J$. The definition $b_{J0} = b_{J1} = \cdots = b_{JK} = 0$ allows the equations for $P_j$ ($j = 1, 2, \ldots, J$) to be written in the compact one-line form in (6.15).

One interprets and tests this general model in the same way as explained for the simpler examples elaborated earlier in Sections 6.1 and 6.2.

## 6.4. RECONCEPTUALIZING CONTRACEPTIVE METHOD CHOICE AS A TWO-STEP PROCESS

The multinomial logit models considered above conceptualize contraceptive method choice as a one-step process. A woman perceives three options: sterilization, some other method, or no method. She considers each of the options and then chooses one of them. The act of choosing among the three options is done all in one step.

An alternative conceptualization is that contraceptive method choice is a two-step process: First a woman chooses whether or not to use contraception at all. If the outcome is to use contraception, the second step is to choose a particular method.

The first of these two steps is appropriately modeled by a binary logit model. The second step, if numerous contraceptive methods are considered, is appropriately modeled by a multinomial logit model. In our previous examples from the Fiji survey, however, only two options (sterilization or "other method") remain once the choice to use contraception is made. Thus, in these examples, the second step is also appropriately modeled by a binary logit model.

In the binary logit model of the first of these two steps, let us define the observed value of the response variable to be 1 if the woman chooses to use contraception (without having yet decided which method) and 0 otherwise. Let us denote the predicted value of this response variable by $P$, the probability of using contraception.

The second binary logit model, pertaining to the second step of choosing a particular method of contraception, is fitted only to the subset of women who actually choose during the first step to use contraception (i.e., to the subset of women whose observed value of the response variable in the first step is 1). In our model of the second step, let us define the observed value of the response variable to be 1 if the woman chooses sterilization and 0 otherwise (otherwise now being "other method"). Let us denote the predicted value of the response variable in this second model by $P'$, the probability of choosing sterilization. $P'$ is a conditional probability. It is conditional on first choosing to use contraception irrespective of method.

Taken together, the results of the two models for the first and second steps imply that the probability that a woman in the original

**TABLE 6.6. Combined MCA Results for the Two-Step Choice Model: Fecund, Nonpregnant, Currently Married Women Aged 15–49 in the 1974 Fiji Fertility Survey**[a]

| Predictor Variable | $PP'$ (Sterilization) | $P(1 - P')$ (Other Method) | $1 - P$ (No Method) |
|---|---|---|---|
| Age | | | |
| 15 | 0 | 18 | 82 |
| 25 | 5 | 40 | 55 |
| 35 | 32 | 29 | 39 |
| 45 | 37 | 25 | 38 |
| Education | | | |
| Low | 16 | 32 | 53 |
| Medium | 15 | 39 | 46 |
| High | 10 | 45 | 44 |
| Residence × ethnicity | | | |
| Urban Indian | 19 | 47 | 33 |
| Urban Fijian | 13 | 27 | 60 |
| Rural Indian | 20 | 41 | 39 |
| Rural Fijian | 5 | 28 | 67 |

[a]The only difference between this table and Table 6.5 is that Table 6.5 assumes a one-step process of contraceptive method choice, whereas this table assumes a two-step process. The models used to generate the two tables have the same variables and were applied to the same input data. See text for further explanation.

sample chooses to use contraception and then chooses sterilization is $PP'$, the probability that she chooses to use contraception and then chooses some other method is $P(1 - P')$, and the probability that she chooses not to use contraception at all is $1 - P$. It is easily verified that these three probabilities add to one, as they must.

How do results from the two-step approach compare with our earlier results from the one-step approach? To find out, we redid Table 6.5 using the two-step approach. We first produced two intermediate MCA tables (not shown), one for each of the two binary logit models corresponding to each of the two steps. Thus the first of these two intermediate tables contained estimates of $P$, and the second contained estimates of $P'$, tabulated in each case by the same predictor variables shown in the row labels of Table 6.5. The probabilities $P$ and $P'$ in these two tables were then combined in the manner described in the previous paragraph. The combined MCA results are shown in Table 6.6, which may be compared with the results in Table 6.5 from the one-step approach. The two tables differ, but not by much. In this example, at least, whether choice is conceptualized as a one-step or a two-step process does not make much difference in the results.

Which model should one use, the one-step model or the two-step model? This question must be answered on the basis of theory and evidence from previous studies. In the case of contraceptive use in Fiji in 1974, however, the choice between the two models is not clear. It seems plausible that many women decided in two steps, in which case a two-step model seems justified. But for others the decision whether to use contraception at all may have been affected by what methods were locally available, in which case a one-step model seems justified. Without more evidence, we cannot say which mode of choice behavior predominates. In the absence of such evidence, it seems prudent to estimate both models and then compare the results.

In a country like the United States, where a wide variety of contraceptive methods are universally available, the decision whether to use contraception at all is not affected by availability of methods, so that a two-step model is more clearly appropriate.

## 6.5. FURTHER READING

More advanced treatments may be found in Theil (1969), Cragg and Uhler (1970), Schmidt and Strauss (1975), Amemiya (1981), and Maddala (1983).