Journal Title: applied regression analysis and generalized linear models

Volume:
Issue:
Month/Year: 2008
Pages: 335-378

Article Author: John Fox

Article Title: Chapter 14 Logit and Probit Models for Categorical Response Variables

CUSTOMER HAS REQUESTED:
Electronic Delivery: Yes
Alternate Delivery Method:
YesHold for pickup

Jason Geller (jg9120)
101 Lassen Court Apt 9
Princeton, NJ 08904

Note

14 Logit and Probit Models for Categorical Response Variables

This chapter and the next deal with generalized linear models—the extension of linear models to variables that have specific non-normal conditional distributions:

- Rather than dive directly into generalized linear models in their full generality, the current chapter takes up linear logit and probit models for categorical response variables. Beginning with this most-important special case allows for a gentler introduction to the topic, I believe. As well, I develop some models for categorical data that are not subsumed by the generalized linear model described in the next chapter.
- Chapter 15 is devoted to the generalized linear model, which has as special cases the linear models of Part II of the text and the dichotomous logit and probit models of the current chapter. Chapter 15 focuses on generalized linear models for count data and develops diagnostic methods for generalized linear models that parallel many of the diagnostics for linear models fit by least-squares, introduced in Part III.

All the statistical models described in previous chapters are for quantitative response variables. It is unnecessary to document the prevalence of qualitative/categorical data in the social sciences. In developing the general linear model, I introduced qualitative *explanatory* variables through the device of coding dummy-variable regressors.[1] There is no reason that qualitative variables should not also appear as response variables, affected by other variables, both qualitative and quantitative.

This chapter deals primarily with logit models for qualitative and ordered-categorical response variables, although related probit models are also briefly considered. The first section of the chapter describes logit and probit models for dichotomous response variables. The second section develops similar statistical models for polytomous response variables, including ordered categories. The third and final section discusses the application of logit models to contingency tables, where the explanatory variables, as well as the response, are categorical.

14.1 Models for Dichotomous Data

Logit and probit models express a qualitative response variable as a function of several explanatory variables, much in the manner of the general linear model. To understand why these models are required, let us begin by examining a representative problem, attempting to apply linear regression to it. The difficulties that are encountered point the way to more satisfactory statistical models for qualitative data.

[1] See Chapter 7.

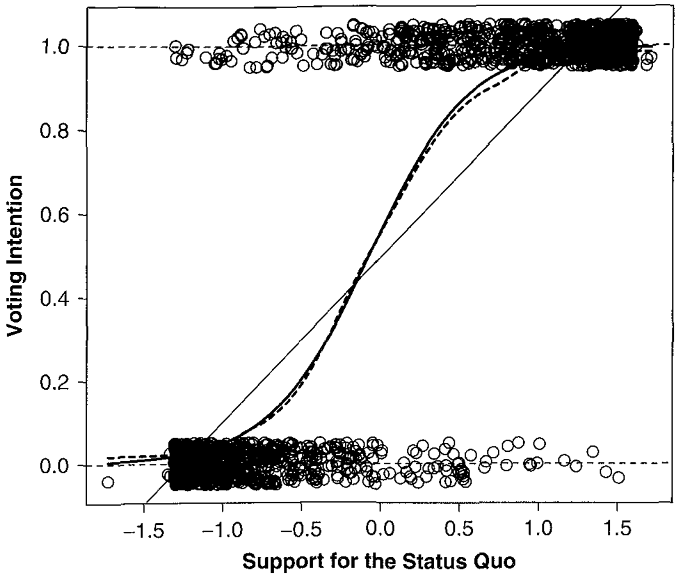

Figure 14.1 Scatterplot of voting intention (1 represents *yes*, 0 represents *no*) by a scale of support for the status quo, for a sample of Chilean voters surveyed prior to the 1988 plebiscite. The points are jittered vertically to minimize overplotting. The solid straight line shows the linear least-squares fit; the solid curved line shows the fit of the logistic-regression model (described in the next section); the broken line represents a nonparametric kernel regression with a span of 0.4.

In September 1988, 15 years after the coup of 1973, the people of Chile voted in a plebiscite to decide the future of the military government headed by General Augusto Pinochet. A *yes* vote would yield eight more years of military rule; a *no* vote would set in motion a process to return the country to civilian government. Of course, the *no* side won the plebiscite, by a clear if not overwhelming margin.

Six months before the plebiscite, the independent research center FLACSO/Chile conducted a national survey of 2,700 randomly selected Chilean voters.[2] Of these individuals, 868 said that they were planning to vote *yes*, and 889 said that they were planning to vote *no*. Of the remainder, 558 said that they were undecided, 187 said that they planned to abstain, and 168 did not answer the question. I will look here only at those who expressed a preference.[3]

Figure 14.1 plots voting intention against a measure of support for the status quo. As seems natural, voting intention appears as a dummy variable, coded 1 for *yes*, 0 for *no*. As we will see presently, this coding makes sense in the context of a dichotomous response variable. Because many points would otherwise be overplotted, voting intention is jittered in the graph (although not in the calculations that follow). Support for the status quo is a scale formed from a number of

[2] FLACSO is an acronym for La Facultad Latino-americano des Ciensias Sociales, a respected institution that conducts social research and trains graduate students in several Latin-American countries. During the Chilean military dictatorship, FLACSO/Chile was associated with the opposition to the military government. I worked on the analysis of the survey described here as part of a joint project between FLACSO in Santiago, Chile, and the Centre for Research on Latin America and the Caribbean at York University, Toronto.

[3] It is, of course, difficult to know how to interpret ambiguous responses such as "undecided." It is tempting to infer that respondents were afraid to state their opinions, but there is other evidence from the survey that this is not the case. Few respondents, for example, uniformly refused to answer sensitive political questions, and the survey interviewers reported little resistance to the survey.

questions about political, social, and economic policies: High scores represent general support for the policies of the miliary regime. (For the moment, disregard the lines plotted in this figure.)

We are used to thinking of a regression as a conditional average. Does this interpretation make sense when the response variable is dichotomous? After all, an average between 0 and 1 represents a "score" for the dummy response variable that cannot be realized by any individual. In the population, the conditional average $E(Y|x_i)$ is simply the proportion of 1s among those individuals who share the value x_i for the explanatory variable—the conditional probability π_i of sampling a *yes* in this group; that is,

$$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$

and, thus,

$$E(Y|x_i) = \pi_i(1) + (1 - \pi_i)(0) = \pi_i \tag{14.1}$$

If X is discrete, then in a sample we can calculate the conditional proportion for Y at each value of X. The collection of these conditional proportions represents the sample nonparametric regression of the dichotomous Y on X. In the present example, X is continuous, but we can nevertheless resort to strategies such as local averaging, as illustrated in Figure 14.1.[4] At low levels of support for the status quo, the conditional proportion of *yes* responses is close to 0; at high levels, it is close to 1; and in between, the nonparametric regression curve smoothly approaches 0 and 1 in a gentle S-shaped pattern.

14.1.1 The Linear-Probability Model

Although nonparametric regression works here, it would be useful to capture the dependency of Y on X as a simple function. To do so will be especially helpful when we introduce additional explanatory variables. As a first effort, let us try linear regression with the usual assumptions:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{14.2}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and ε_i and ε_j are independent for $i \neq j$. If X is random, then we assume that it is independent of ε.

Under Equation 14.2, $E(Y_i) = \alpha + \beta X_i$, and so, from Equation 14.1,

$$\pi_i = \alpha + \beta X_i$$

For this reason, the linear-regression model applied to a dummy response variable is called the *linear-probability model.* This model is untenable, but its failure will point the way toward more adequate specifications:

- Because Y_i can take on only the values 0 and 1, the error ε_i is dichotomous as well—and, hence, is not normally distributed, as assumed: If $Y_i = 1$, which occurs with probability π_i, then

$$\varepsilon_i = 1 - E(Y_i) = 1 - (\alpha + \beta X_i) = 1 - \pi_i$$

[4] The nonparametric-regression line in Figure 14.1 was fit by kernel regression—a method based on locally weighted averaging, which is similar to locally weighted regression (lowess, which was introduced in Chapter 2 for smoothing scatterplots). Unlike lowess, however, the kernel estimator of a proportion cannot be outside the interval from 0 to 1. Both the kernel-regression estimator and other nonparametric-regression methods that are more appropriate for a dichotomous response are described in Chapter 18. The span for the kernel regression (i.e., the fraction of the data included in each local average) is 0.4.

Alternatively, if $Y_i = 0$, which occurs with probability $1 - \pi_i$, then

$$\varepsilon_i = 0 - E(Y_i) = 0 - (\alpha + \beta X_i) = 0 - \pi_i = -\pi_i$$

Because of the central-limit theorem, however, the assumption of normality is not critical to least-squares estimation of the normal-probability model, as long as the sample size is sufficiently large.

- The variance of ε cannot be constant, as we can readily demonstrate: If the assumption of linearity holds over the range of the data, then $E(\varepsilon_i) = 0$. Using the relations just noted,

  $$V(\varepsilon_i) = \pi_i(1 - \pi_i)^2 + (1 - \pi_i)(-\pi_i)^2 = \pi_i(1 - \pi_i)$$

  The heteroscedasticity of the errors bodes ill for ordinary-least-squares estimation of the linear probability model, but only if the probabilities π_i get close to 0 or 1.[5] Goldberger (1964, pp. 248–250) has proposed a correction for heteroscedasticity employing weighted least squares.[6] Because the variances $V(\varepsilon_i)$ depend on the π_i, however, which, in turn, are functions of the unknown parameters α and β, we require preliminary estimates of the parameters to define weights. Goldberger obtains ad hoc estimates from a preliminary OLS regression; that is, he takes $\widehat{V}(\varepsilon_i) = \widehat{Y}_i(1 - \widehat{Y}_i)$. The fitted values from an OLS regression are not constrained to the interval [0, 1], and so some of these "variances" may be negative.
- This last remark suggests the most serious problem with the linear-probability model: The assumption that $E(\varepsilon_i) = 0$—that is, the assumption of linearity—is only tenable over a limited range of X values. If the range of the Xs is sufficiently broad, then the linear specification cannot confine π to the unit interval [0, 1]. It makes no sense, of course, to interpret a number outside the unit interval as a probability. This difficulty is illustrated in Figure 14.1, in which the least-squares line fit to the Chilean plebiscite data produces fitted probabilities below 0 at low levels and above 1 at high levels of support for the status quo.

Dummy *regressor* variables do not cause comparable difficulties because the general linear model makes no distributional assumptions about the regressors (other than independence from the errors). Nevertheless, for values of π not too close to 0 or 1, the linear-probability model estimated by least squares frequently provides results similar to those produced by the more generally adequate methods described in the remainder of this chapter.

> **It is problematic to apply least-squares linear regression to a dichotomous response variable: The errors cannot be normally distributed and cannot have constant variance. Even more fundamentally, the linear specification does not confine the probability for the response to the unit interval.**

One solution to the problems of the linear-probability model—though not a good general solution—is simply to constrain π to the unit interval while retaining the linear relationship between π and X within this interval:

$$\pi = \begin{cases} 0 & \text{for } 0 > \alpha + \beta X \\ \alpha + \beta X & \text{for } 0 \le \alpha + \beta X \le 1 \\ 1 & \text{for } \alpha + \beta X > 1 \end{cases} \tag{14.3}$$

[5] See Exercise 14.1. Remember, however, that it is the *conditional probability*, not the *marginal probability*, of Y that is at issue: The overall proportion of 1s can be near .5 (as in the Chilean plebiscite data), and yet the conditional proportion can still get very close to 0 or 1.

[6] See Section 12.2.2 for a discussion of weighted-least-squares estimation.

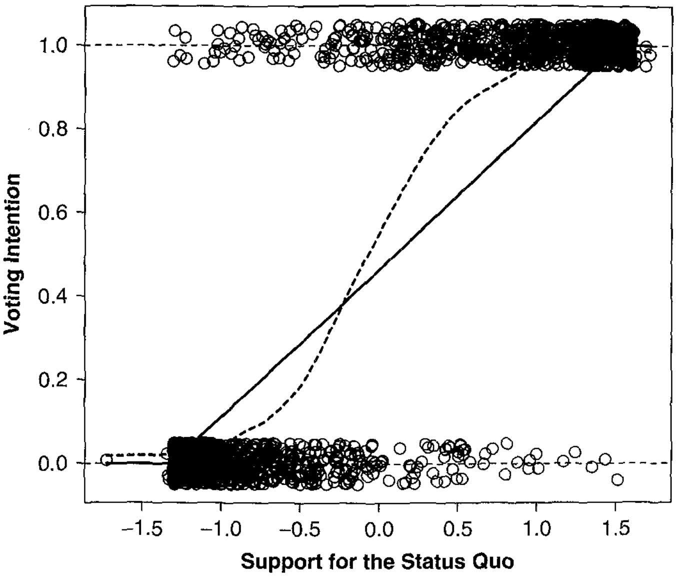

Figure 14.2 The solid line shows the constrained linear-probability model, fit by maximum likelihood to the Chilean plebiscite data. The broken line is for a nonparametric kernel regression with a span of 0.4.

Figure 14.2 shows the fit of this model to the Chilean plebiscite data, with the parameters α and β estimated by maximum likelihood. Although this *constrained linear-probability model* cannot be dismissed on logical grounds, the model has certain unattractive features: Most importantly, the abrupt changes in slope at $\pi = 0$ and $\pi = 1$ are usually unreasonable. A smoother relationship between π and X (as characterizes the nonparametric regression in Figure 14.1) is more generally sensible. Moreover, numerical instability can make the constrained linear-probability model difficult to fit to data, and the statistical properties of estimators of the model are hard to derive because of the discontinuities in the slope.[7]

14.1.2 Transformations of π: Logit and Probit Models

A central difficulty of the unconstrained linear-probability model is its inability to ensure that π stays between 0 and 1. What we require to correct this problem is a positive monotone (i.e., nondecreasing) function that maps the *linear predictor* $\eta = \alpha + \beta X$ into the unit interval. A transformation of this type will allow us to retain the fundamentally linear structure of the model while avoiding the contradiction of probabilities below 0 or above 1. Any cumulative probability distribution function (CDF) meets this requirement.[8] That is, we can respecify the model as

$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i) \tag{14.4}$$

[7] *Consider the strong constraints that the data place on the maximum-likelihood estimators of α and β: If, as in the illustration, $\widehat{\beta} > 0$, then the rightmost observation for which $Y = 0$ can have an X value no larger than $(1 - \widehat{\alpha})/\widehat{\beta}$, which is the point at which the estimated regression line hits $\widehat{\pi} = 1$, because any 0 to the right of this point would produce a 0 likelihood. Similarly, the leftmost observation for which $Y = 1$ can have an X value no smaller than $-\widehat{\alpha}/\widehat{\beta}$, the point at which the regression line hits $\widehat{\pi} = 0$. As the sample size grows, these extreme values will tend to to move, respectively, to the right and left, making $\widehat{\beta}$ smaller.

[8] See Appendix D on probability and estimation.

where the CDF $P(\cdot)$ is selected in advance, and α and β are then parameters to be estimated.

If we choose $P(\cdot)$ as the cumulative rectangular distribution, for example, then we obtain the constrained linear-probability model (Equation 14.3).[9] An a priori reasonable $P(\cdot)$ should be both smooth and symmetric and should approach $\pi = 0$ and $\pi = 1$ as asymptotes.[10] Moreover, it is advantageous if $P(\cdot)$ is strictly increasing, for then the transformation (Equation 14.4) is one to one, permitting us to rewrite the model as

$$P^{-1}(\pi_i) = \eta_i = \alpha + \beta X_i \tag{14.5}$$

where $P^{-1}(\cdot)$ is the inverse of the CDF $P(\cdot)$ (i.e., the quantile function for the distribution).[11] Thus, we have a linear model (Equation 14.5) for a transformation of π, or—equivalently—a nonlinear model (Equation 14.4) for π itself.

The transformation $P(\cdot)$ is often chosen as the CDF of the unit-normal distribution, $N(0, 1)$,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}Z^2\right) dZ \tag{14.6}$$

or, even more commonly, of the *logistic distribution*

$$\Lambda(z) = \frac{1}{1+e^{-z}} \tag{14.7}$$

In these equations, $\pi \approx 3.141$ and $e \approx 2.718$ are the familiar mathematical constants.[12]

- Using the normal distribution $\Phi(\cdot)$ yields the *linear probit model*:

$$\begin{aligned} \pi_i &= \Phi(\alpha + \beta X_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} \exp\left(-\frac{1}{2}Z^2\right) dZ \end{aligned}$$

- Using the logistic distribution $\Lambda(\cdot)$ produces the *linear logistic-regression* or *linear logit model*:

$$\begin{aligned} \pi_i &= \Lambda(\alpha + \beta X_i) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta X_i)]} \end{aligned} \tag{14.8}$$

Once their variances are equated—the logistic distribution has variance $\pi^2/3$—the logit and probit transformations are so similar that it is not possible, in practice, to distinguish between them without a great deal of data, as is apparent in Figure 14.3. It is also clear from this graph that both functions are nearly linear over much of their range, say between about $\pi = .2$ and $\pi = .8$. This is why the linear-probability model produces results similar to the logit and probit models, except for extreme values of π_i.

[9] See Exercise 14.2.

[10] This is not to say, however, that $P(\cdot)$ needs to be symmetric in every case, just that symmetric $P(\cdot)$s are more appropriate *in general*. For an example of an asymmetric choice of $P(\cdot)$, see the discussion of the complementary log-log transformation in Chapter 15.

[11] If, alternatively, the CDF levels off (as is the case, e.g., for the rectangular distribution), then the inverse of the CDF does not exist.

[12] A note to the reader for whom calculus is unfamiliar: An integral, represented by the symbol $\int$ in Equation 14.6, represents the area under a curve, here the area between $Z = -\infty$ and $Z = z$ under the curve given by the function $\exp\left(-\frac{1}{2}Z^2\right)$. The constant $1/\sqrt{2\pi}$ insures that the total area under the normal density function "integrates" (i.e., adds up) to 1.

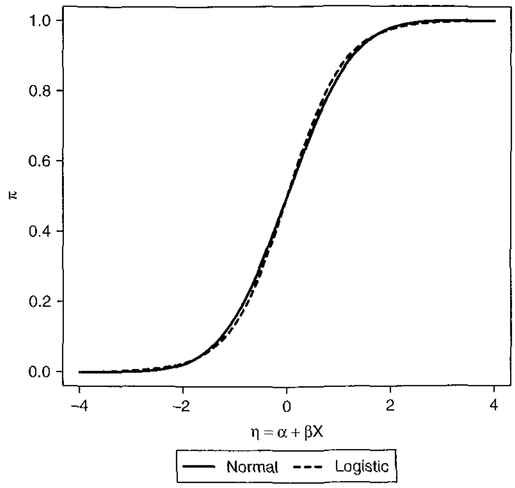

Figure 14.3 Once their variances are equated, the cumulative logistic and cumulative normal distributions—used here to transform $\eta = \alpha + \beta X$ to the unit interval—are virtually indistinguishable.

Despite their essential similarity, there are two practical advantages of the logit model compared to the probit model:

1. The equation of the logistic CDF (Equation 14.7) is very simple, while the normal CDF (Equation 14.6) involves an unevaluated integral. This difference is trivial for dichotomous data because very good closed-form approximations to the normal CDF are available, but for polytómous data, where we will require the *multivariate* logistic or normal distribution, the disadvantage of the probit model is somewhat more acute.[13]

2. The inverse linearizing transformation for the logit model, $\Lambda^{-1}(\pi)$, is directly interpretable as a *log-odds*, while the inverse transformation for the probit model, the quantile function of the standard-normal distribution, $\Phi^{-1}(\pi)$, does not have a direct interpretation. Rearranging Equation 14.8, we get

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \beta X_i) \tag{14.9}$$

The ratio $\pi_i/(1-\pi_i)$ is the *odds* that $Y_i = 1$ (e.g., the odds of voting *yes*), an expression of relative chances familiar from gambling (at least to those who engage in this vice). Unlike the probability scale, odds are unbounded above (though bounded below by 0). Taking the log of both sides of Equation 14.9 produces

$$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$

The inverse transformation $\Lambda^{-1}(\pi) = \log_e[\pi/(1-\pi)]$, called the *logit* of π, is therefore the log of the odds that Y is 1 rather than 0. As the following table shows, if the odds are "even"—that is, equal to 1, corresponding to $\pi = .5$—then the logit is 0. The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response variable in a linear-like model.

[13] See Section 14.2.1.

Probability π	*Odds* $\dfrac{\pi}{1-\pi}$	*Logit* $\log_e \dfrac{\pi}{1-\pi}$
.01	1/99 = 0.0101	−4.60
.05	5/95 = 0.0526	−2.94
.10	1/9 = 0.1111	−2.20
.30	3/7 = 0.4286	−0.85
.50	5/5 = 1	0.00
.70	7/3 = 2.3333	0.85
.90	9/1 = 9	2.20
.95	95/5 = 19	2.94
.99	99/1 = 99	4.60

The logit model is a linear, additive model for the log odds, but (from Equation 14.9) it is also a multiplicative model for the odds:

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \beta X_i) = \exp(\alpha)\exp(\beta X_i)$$
$$= e^{\alpha}\left(e^{\beta}\right)^{X_i}$$

So, increasing X by 1 changes the logit by β and multiplies the odds by e^{β}. For example, if $\beta = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$.[14]

Still another way of understanding the parameter β in the logit model is to consider the slope of the relationship between π and X, given by Equation 14.8. Because this relationship is nonlinear, the slope is not constant; the slope is $\beta\pi(1-\pi)$ and hence is at a maximum when $\pi = \frac{1}{2}$, where the slope is $\beta\frac{1}{2}(1-\frac{1}{2}) = \beta/4$, as illustrated in the following table:[15]

π	$\beta\pi(1-\pi)$
.01	$\beta\times.0099$
.05	$\beta\times.0475$
.10	$\beta\times.09$
.20	$\beta\times.16$
.50	$\beta\times.25$
.80	$\beta\times.16$
.90	$\beta\times.09$
.95	$\beta\times.0475$
.99	$\beta\times.0099$

Notice that the slope of the relationship between π and X does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

The least-squares line fit to the Chilean plebiscite data in Figure 14.1, for example, has the equation

$$\widehat{\pi}_{\text{yes}} = 0.492 + 0.394 \times \text{Status quo} \tag{14.10}$$

[14] The exponentiated coefficient e^{β} is sometimes called an "odds ratio" because it represents the ratio of the odds of response at two X values, with the X value in the numerator one unit larger than that in the denominator.

[15] See Exercise 14.3.

As I have pointed out, this line is a poor summary of the data. The logistic-regression model, fit by the method of maximum likelihood (to be developed presently), has the equation

$$\log_e \frac{\widehat{\pi}_{\text{yes}}}{\widehat{\pi}_{\text{no}}} = 0.215 + 3.21 \times \text{Status quo}$$

As is apparent from Figure 14.1, the logit model produces a much more adequate summary of the data, one that is very close to the nonparametric regression. Increasing support for the status quo by one unit multiplies the odds of voting *yes* by $e^{3.21} = 24.8$. Put alternatively, the slope of the relationship between the fitted probability of voting *yes* and support for the status quo at $\widehat{\pi}_{\text{yes}} = .5$ is $3.21/4 = 0.80$. Compare this value with the slope ($B = 0.39$) from the linear least-squares regression in Equation 14.10.[16]

14.1.3 An Unobserved-Variable Formulation

An alternative derivation of the logit or probit model posits an underlying regression for a continuous but unobservable response variable ξ (representing, e.g., the "propensity" to vote *yes*), scaled so that

$$Y_i = \begin{cases} 0 & \text{when } \xi_i \leq 0 \\ 1 & \text{when } \xi_i > 0 \end{cases} \tag{14.11}$$

That is, when ξ crosses 0, the observed discrete response Y changes from *no* to *yes*. The latent variable ξ is assumed to be a linear function of the explanatory variable X and the (usual) unobservable error variable ε:

$$\xi_i = \alpha + \beta X_i - \varepsilon_i \tag{14.12}$$

(It is notationally convenient here—but otherwise inconsequential—to *subtract* the error ε.) We want to estimate the parameters α and β, but cannot proceed by least-squares regression of ξ on X because the latent response variable (unlike Y) is not observed.

Using Equations 14.11 and 14.12,

$$\begin{aligned} \pi_i \equiv \Pr(Y_i = 1) &= \Pr(\xi_i > 0) = \Pr(\alpha + \beta X_i - \varepsilon_i > 0) \\ &= \Pr(\varepsilon_i < \alpha + \beta X_i) \end{aligned}$$

If the errors are independently distributed according to the unit-normal distribution, $\varepsilon_i \sim N(0, 1)$, then

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Phi(\alpha + \beta X_i)$$

which is the probit model.[17] Alternatively, if the ε_i follow the similar logistic distribution, then we get the logit model

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Lambda(\alpha + \beta X_i)$$

[16] As I have explained, the slope for the logit model is not constant: It is steepest at $\pi = .5$ and flattens out as π approaches 0 and 1. The linear probability model, therefore, will agree more closely with the logit model when the response probabilities do not (as here) attain extreme values.

[17] The variance of the errors is set conveniently to 1. This choice is legitimate because we have not yet fixed the unit of measurement of the latent variable ξ. The location of the ξ scale was implicitly fixed by setting 0 as the point at which the observable response changes from *no* to *yes*. You may be uncomfortable assuming that the errors for an unobservable response variable are normally distributed, because we cannot check the assumption by examining residuals, for example. In fact, in most instances we can ensure that the error distribution has any form we please by transforming ξ to make the assumption true. We cannot, however, simultaneously ensure that the true regression is linear. If the latent-variable regression is not linear, then the probit model will not adequately capture the relationship between the dichotomous Y and X.

We will have occasion to return to the unobserved-variable formulation of logit and probit models when we consider models for ordinal categorical data.[18]

14.1.4 Logit and Probit Models for Multiple Regression

Generalizing the logit and probit models to several explanatory variables is straightforward. All we require is a linear predictor (η_i in Equation 14.13) that is a function of several regressors. For the logit model,

$$\begin{aligned}\pi_i &= \Lambda(\eta_i) = \Lambda(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})]}\end{aligned} \tag{14.13}$$

or, equivalently,

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

For the probit model,

$$\pi_i = \Phi(\eta_i) = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

Moreover, the Xs can be as general as in the general linear model, including, for example:

- quantitative explanatory variables;
- transformations of quantitative explanatory variables;
- polynomial regressors formed from quantitative explanatory variables;
- dummy regressors representing qualitative explanatory variables; and
- interaction regressors.

Interpretation of the partial-regression coefficients in the general linear logit model (Equation 14.13) is similar to the interpretation of the slope in the logit simple-regression model, with the additional provision of holding other explanatory variables in the model constant. For example, expressing the model in terms of odds,

$$\begin{aligned}\frac{\pi_i}{1 - \pi_i} &= \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}) \\ &= e^{\alpha} \left(e^{\beta_1}\right)^{X_{i1}} \cdots \left(e^{\beta_k}\right)^{X_{ik}}\end{aligned}$$

Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other Xs constant. Similarly, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

> More adequate specifications than the linear probability model transform the linear predictor $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ smoothly to the unit interval, using a cumulative probability distribution function $P(\cdot)$. Two such specifications are the probit and the logit models, which use the normal and logistic CDFs, respectively. Although these models are very similar, the logit model is simpler to interpret because it can be written as a linear model for the log odds: $\log_e[\pi_i/(1 - \pi_i)] = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$.

[18] See Section 14.2.3.

The general linear logit and probit models can be fit to data by the method of maximum likelihood. I will concentrate here on outlining maximum-likelihood estimation for the logit model. Details are given in the next section.

Recall that the response variable Y_i takes on the two values 1 and 0 with probabilities π_i and $1 - \pi_i$, respectively. Using a mathematical "trick," the probability distribution for Y_i can be compactly represented as a single equation:[19]

$$p(y_i) \equiv \Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

where y_i can be 0 or 1.

Now consider a particular sample of n independent observations, $y_1, y_2, \ldots, y_n$ (comprising a specific sequence of 0s and 1s). Because the observations are independent, the joint probability for the data is the product of the marginal probabilities:

$$\begin{aligned} p(y_1, y_2, \ldots, y_n) &= p(y_1)p(y_2)\cdots p(y_n) \qquad (14.14)\\ &= \prod_{i=1}^{n} p(y_i)\\ &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}\\ &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i}(1 - \pi_i) \end{aligned}$$

From the general logit model (Equation 14.13),

$$\frac{\pi_i}{1 - \pi_i} = \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

and (after some manipulation)[20]

$$1 - \pi_i = \frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}$$

Substituting these results into Equation 14.14 expresses the probability of the data in terms of the parameters of the logit model:

$$\begin{aligned} &p(y_1, y_2, \ldots, y_n)\\ &\quad = \prod_{i=1}^{n}[\exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})]^{y_i}\left(\frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}\right) \end{aligned}$$

Thinking of this equation as a function of the parameters, and treating the data $(y_1, y_2, \ldots, y_n)$ as fixed, produces the likelihood function, $L(\alpha, \beta_1, \ldots, \beta_k)$. The values of $\alpha, \beta_1, \ldots, \beta_k$ that maximize $L(\alpha, \beta_1, \ldots, \beta_k)$ are the maximum-likelihood estimates $A, B_1, \ldots, B_k$.

Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.[21] For an individual coefficient, it is most convenient to test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$ by calculating the Wald statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\mathrm{SE}(B_j)}$$

[19] See Exercise 14.4.

[20] See Exercise 14.5.

[21] These general procedures are discussed in Appendix D on probability and estimation.

where $\mathrm{SE}(B_j)$ is the asymptotic (i.e., large-sample) standard error of B_j. To test the most common hypothesis, H_0: $\beta_j = 0$, we simply divide the estimated coefficient by its standard error to compute $Z_0 = B_j/\mathrm{SE}(B_j)$; these tests are analogous to t-tests for individual coefficients in the general linear model. The test statistic Z_0 follows an asymptotic standard-normal distribution under the null hypothesis, an approximation that is usually reasonably accurate unless the sample size is small.[22] Similarly, an asymptotic $100(1-a)\%$ confidence interval for β_j is given by

$$\beta_j = B_j \pm z_{a/2}\mathrm{SE}(B_j)$$

where $z_{a/2}$ is the value from $Z \sim N(0, 1)$ with probability $a/2$ to the right. Wald tests and joint confidence regions for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.[23]

It is also possible to formulate a likelihood-ratio test for the hypothesis that several coefficients are simultaneously 0, H_0: $\beta_1 = \cdots = \beta_q = 0$. We proceed, as in least-squares regression, by fitting two models to the data: the full model (model 1),

$$\operatorname{logit}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k$$

and the null model (model 0),

$$\begin{aligned}\operatorname{logit}(\pi) &= \alpha + 0X_1 + \cdots + 0X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k \\ &= \alpha + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k\end{aligned}$$

Fitting each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model. Because the null model is a specialization of the full model, $L_1 \geq L_0$. The generalized likelihood-ratio test statistic for the null hypothesis is

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

Under the null hypothesis, this test statistic has an asymptotic chi-square distribution with q degrees of freedom.

By extension, a test of the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ is obtained by specifying a null model that includes only the regression constant, $\operatorname{logit}(\pi) = \alpha$. At the other extreme, the likelihood-ratio test can, of course, be applied to a *single* coefficient, H_0: $\beta_j = 0$, and this test can be inverted to provide a confidence interval for β_j: For example, the 95% confidence interval for β_j includes all values β'_j for which the hypothesis H_0: $\beta_j = \beta'_j$ is acceptable at the .05 level—that is, all values of β'_j for which $2(\log_e L_1 - \log_e L_0) \leq \chi^2_{.05,1} = 3.84$, where $\log_e L_1$ is (as before) the maximized log likelihood for the full model, and $\log_e L_0$ is the maximized log likelihood for a model in which β_j is constrained to the value β'_j.

An analog to the multiple-correlation coefficient can also be obtained from the log likelihood. The maximized log likelihood for the fitted model can be written as[24]

$$\log_e L = \sum_{i=1}^{n} \left[y_i \log_e P_i + (1 - y_i) \log_e (1 - P_i) \right]$$

[22] Under certain circumstances, however, tests and confidence intervals based on the Wald statistic can break down in logistic regression (see Hauck & Donner, 1977). Tests and confidence intervals based on the likelihood-ratio statistic, described immediately below, are more reliable, though more time-consuming to compute.

[23] See Section 14.1.5.

[24] See Exercise 14.6.

where P_i is the fitted probability that $Y_i = 1$,[25] that is,

$$P_i = \frac{1}{1 + \exp[-(A + B_1 X_{i1} + \cdots + B_k X_{ik})]}$$

Thus, if the fitted model can perfectly predict the Y values ($P_i = 1$ whenever $y_i = 1$, and $P_i = 0$ whenever $y_i = 0$), then $\log_e L = 0$ (i.e., the maximized likelihood is $L = 1$).[26] To the extent that predictions are less than perfect, $\log_e L < 0$ (and $0 < L < 1$).

By comparing $\log_e L_0$ for the model containing only the constant to $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y. The quantity $G^2 \equiv -2 \log_e L$, called the *residual deviance* under the model, is a generalization of the residual sum of squares for a linear model.[27] Thus,

$$\begin{aligned} R^2 &\equiv 1 - \frac{G_1^2}{G_0^2} \\ &= 1 - \frac{\log_e L_1}{\log_e L_0} \end{aligned}$$

is analogous to R^2 for a linear model.[28]

> The dichotomous logit model can be fit to data by the method of maximum likelihood. Wald tests and likelihood-ratio tests for the coefficients of the model parallel t-tests and incremental F-tests for the general linear model. The deviance for the model, defined as $G^2 = -2 \times$ the maximized log likelihood, is analogous to the residual sum of squares for a linear model.

To illustrate logistic regression, I turn once again to the 1994 wave of the Statistics Canada Survey of Labour and Income Dynamics (the "SLID").[29] Confining our attention to married women between the ages of 20 and 35, I examine how the labor-force participation of these women (defined as working outside the home at some point during the year of the survey) is related to several explanatory variables:

- the region of the country in which the woman resides;
- the presence of children between 0 and 4 years of age in the household, coded as absent or present;
- the presence of children between 5 and 9 years of age;
- the presence of children between 10 and 14 years of age;
- family after-tax income, excluding the woman's own income (if any);[30] and
- education, defined as number of years of schooling.

[25] Of course, in a particular sample, y_i is either 0 or 1, so we can interpret this fitted probability as the estimated population proportion of individuals sharing the ith person's characteristics for whom Y is 1. Other interpretations are also possible, but this is the most straightforward.

[26] Because, for the logit model, π never quite reaches 0 or 1, the predictions cannot be perfect, but they can approach perfection in the limit.

[27] See Exercise 14.7 and Chapter 15 on generalized linear models.

[28] For alternative R^2 measures for logit and probit models, see, for example, Veall and Zimmermann (1996).

[29] The SLID was introduced in Chapter 2.

[30] I excluded from the analysis two women for whom this variable is negative.

Table 14.1 Distributions of Variables in the SLID Data Set

Variable	*Summary*
Labor-Force Participation	Yes, 79%
Region (R)	Atlantic, 23%; Quebec, 13; Ontario, 30; Prairies, 26; BC, 8
Children 0–4 (K04)	Yes, 53%
Children 5–9 (K59)	Yes, 44%
Children 10–14 (K1014)	Yes, 22%
Family Income (I, $1,000s)	5-number summary: 0, 18.6, 26.7, 35.1, 131.1
Education (E, years)	5-number summary: 0, 12, 13, 15, 20

The SLID data set includes 1,936 women with valid data on these variables. Some information about the distribution of the variables appears in Table 14.1. Recall that the five-number summary includes the minimum, first quartile, median, third quartile, and maximum of a variable.

In modeling these data, I want to allow for the possibility of interaction between presence of children and each of family income and education in determining women's labor-force participation. Table 14.2 shows the residual deviances and number of parameters for each of a series of models fit to the SLID data. These models are formulated so that likelihood-ratio tests of terms in the full model can be computed by taking differences in the residual deviances for the models, in conformity with the principle of marginality. The residual deviances are the building blocks of likelihood-ratio tests, much as residual sums of squares are the building blocks of incremental F-tests in linear models. The tests themselves, with an indication of the models contrasted for each test, appear in an *analysis-of-deviance* table in Table 14.3, closely analogous to an ANOVA table for a linear model.

It is clear from the likelihood-ratio tests in Table 14.3 that none of the interactions approach statistical significance. Presence of children four years old and younger and education have very highly statistically significant coefficients; the terms for region, children 5 to 9 years old, and family income are also statistically significant, while that for children 10 through 14 is not.

Estimated coefficients and standard errors for a summary model including the statistically significant terms are given in Table 14.4. The Atlantic provinces are the baseline category for the region effects in this model. The column of the table labelled e^{B_j} represents multiplicative effects on the odds scale. Thus, for example, holding the other explanatory variables constant, having children 0 to 4 years old in the household reduces the *odds* of labor-force participation by $100(1 - 0.379) = 62.1\%$; and increasing education by 1 year increases the odds of labor-force participation by $100(1.246 - 1) = 24.6\%$. As explained, as long as a coefficient is not too large, we can also express effects on the probability scale near $\pi = .5$ by dividing the coefficient by 4: For example (and again, holding other explanatory variables constant), if the probability of labor-force participation is near .5 with children 0 to 4 absent, the presence of children of this age decreases the probability by approximately $0.9702/4 = 0.243$ or 24.3%, while an additional year of education increases the probability by approximately $0.2197/4 = .0549$ or 5.5%.

Still another strategy for interpreting a logit model is to graph the high-order terms in the model, producing effect displays, much as we did for linear models.[31] The final model for the SLID labor-force participation data in Table 14.4 has a simple structure in that there are no interactions or polynomial terms. Nevertheless, it helps to see how each explanatory variable influences the probability of the response holding other explanatory variables to their average

[31] See the discussion of effect displays for linear models in Sections 7.3.4 and 8.3.2. Details of effect displays for logit models are developed in a more general context in the next chapter (Section 15.3.4).

Table 14.2 Models Fit to the SLID Labor-Force Participation Data

Model	*Terms in the Model*	*Number of Parameters*	*Residual Deviance*
0	C	1	1988.084
1	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K1014×I, K04×E, K59×E, K1014×E	16	1807.376
2	Model 1 – K04×I	15	1807.378
3	Model 1 – K59×I	15	1808.600
4	Model 1 – K1014×I	15	1807.834
5	Model 1 – K04×E	15	1807.407
6	Model 1 – K59×E	15	1807.734
7	Model 1 – K1014×E	15	1807.938
8	Model 1 – R	12	1824.681
9	C, R, K04, K59, K1014, I, E, K59×I, K1014×I, K59×E, K1014×E	14	1807.408
10	Model 9 – K04	13	1866.689
11	C, R, K04, K59, K1014, I, E, K04×I, K1014×I, K04×E, K1014×E	14	1809.268
12	Model 11 – K59	13	1819.273
13	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K04×E, K59×E	14	1808.310
14	Model 13 – K1014	13	1808.548
15	C, R, K04, K59, K1014, I, E, K04×E, K59×E, K1014×E	13	1808.854
16	Model 15 – I	12	1817.995
17	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K1014×I	13	1808.428
18	Model 17 – E	12	1889.223

NOTE: "C" represents the regression constant; codes for other variables in the model are given in Table 14.1.

Table 14.3 Analysis of Deviance Table for the SLID Labor-Force Participation Logit Model

Term	*Models Contrasted*	*df*	G_0^2	*p*
Region (R)	8-1	4	17.305	.0017
Children 0–4 (K04)	10-9	1	59.281	$\ll$.0001
Children 5–9 (K59)	12-11	1	10.005	.0016
Children 10–14 (K1014)	14-12	1	0.238	.63
Family Income (I)	16-15	1	9.141	.0025
Education (E)	18-17	1	80.795	$\ll$.0001
K04×I	2-1	1	0.002	.97
K59×I	3-1	1	1.224	.29
K1014×I	4-1	1	0.458	.50
K04×E	5-1	1	0.031	.86
K59×E	6-1	1	0.358	.55
K1014×E	7-1	1	0.562	.45

Table 14.4 Estimates for a Final Model Fit to the SLID Labor-Force Participation Data

Coefficient	*Estimate (B_j)*	*Standard Error*	e^{B_j}
Constant	−0.3763	0.3398	
Region: Quebec	−0.5469	0.1899	0.579
Region: Ontario	0.1038	0.1670	1.109
Region: Prairies	0.0742	0.1695	1.077
Region: BC	0.3760	0.2577	1.456
Children 0–4	−0.9702	0.1254	0.379
Children 5–9	−0.3971	0.1187	0.672
Family income ($1,000s)	−0.0127	0.0041	0.987
Education (years)	0.2197	0.0250	1.246
Residual deviance	1810.444		

values. In Figure 14.4, I plot the terms in the model on the logit scale (given by the left-hand axis in each graph), preserving the linear structure of the model, but I also show corresponding fitted probabilities of labor-force participation (on the right-hand axis)—a more familiar scale on which to interpret the results.

We should not forget that the logit model fit to the SLID data is a parametric model, assuming linear partial relationships (on the logit scale) between labor-force participation and the two quantitative explanatory variables, family income and education. There is no more reason to believe that relationships are necessarily linear in logit models than in linear least-squares regression. I will take up diagnostics, including nonlinearity diagnostics, for logit models and other generalized linear models in the next chapter.[32]

Woes of Logistic-Regression Coefficients

Just as the least-squares surface flattens out at its minimum when the Xs are collinear, the likelihood surface for a logistic regression flattens out at its maximum in the presence of collinearity, so that the maximum-likelihood estimates of the coefficients of the model are not uniquely defined. Likewise, strong, but less-than-perfect, collinearity causes the coefficients to be imprecisely estimated.

Paradoxically, problems for estimation can also occur in logit models when the explanatory variables are very strong predictors of the dichotomous response. One such circumstance, illustrated in Figure 14.5, is *separability*. When there is a single X, the data are separable if the "failures" (0s) and "successes" (1s) fail to overlap [as in Figure 14.5(a)]. In this case, the maximum-likelihood estimate of the slope coefficient β is infinite (either $-\infty$ or $+\infty$, depending on the direction of the relationship between X and Y), and the estimate of the intercept α is not unique. When there are two Xs, the data are separable if there is a line in the $\{X_1, X_2\}$ plane that separates successes from failures [as in Figure 14.5(b)]. For three Xs, the data are separable if there is a separating plane in the three-dimensional space of the Xs; and the generalization to any number of Xs is a separating hyperplane—that is, a linear surface of dimension $k-1$ in the k-dimensional X space.

[32] See Section 15.4.

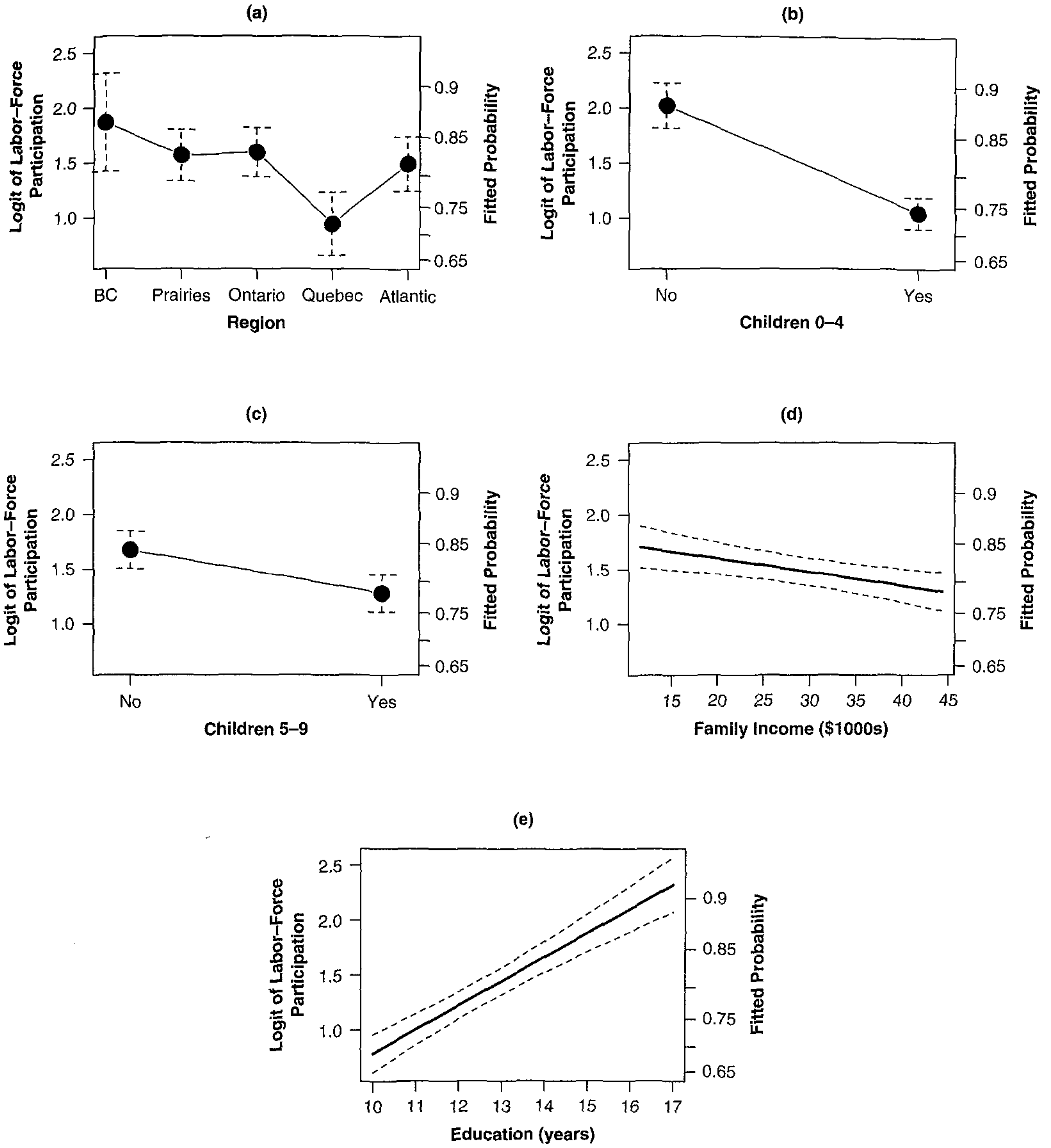

Figure 14.4 Effect displays for the summary logit model fit to the SLID labor-force participation data. The error bars and envelopes give pointwise 95% confidence intervals around the estimated effects. The plots for family income and education range between the 10th and 90th percentiles of these variables.

Still another circumstance that yields infinite coefficients is that of data in which some of the responses become perfectly predictable even in the absence of complete separability. For example, if at one level of a factor all observations are successes, the estimated probability of success for an observation at this level is 1, and the odds of success are $1/0 = \infty$.

Statistical software may or may not detect these problems for estimation. The problems may manifest themselves in failure of the software to converge to a solution, in wildly large estimated coefficients, or in very large coefficient standard errors.

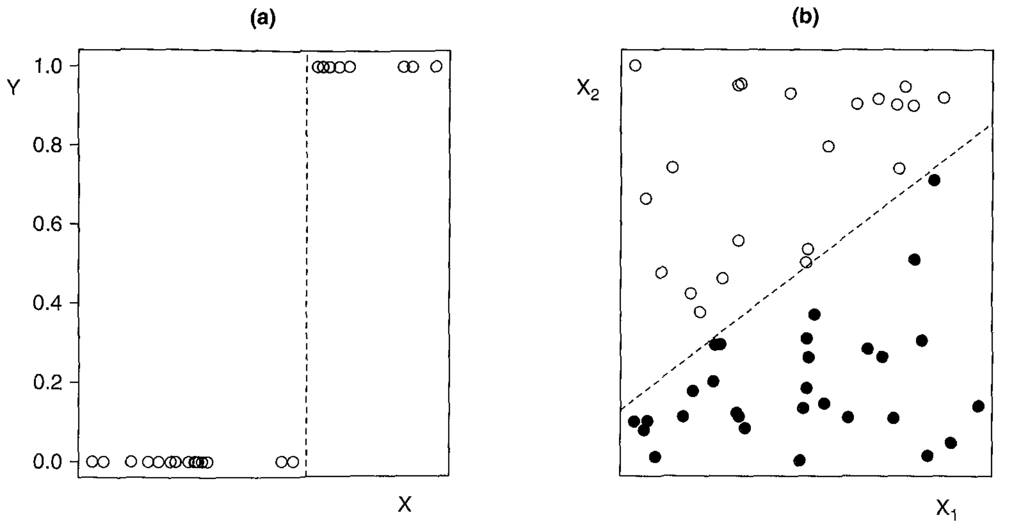

Figure 14.5 Separability in logistic regression: (a) with one explanatory variable, X; (b) with two explanatory variables, X_1 and X_2. In panel *(b)*, the solid dots represent observations for which $Y = 1$, and the hollow dots observations for which $Y = 0$.

14.1.5 Estimating the Linear Logit Model*

In this section, I will develop the details of maximum-likelihood estimation for the general linear logit model (Equation 14.13 on page 344). It is convenient to rewrite the model in vector form as

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{x}_i'\beta)}$$

where $\mathbf{x}_i' \equiv (1, X_{i1}, \ldots, X_{ik})$ is the ith row of the model matrix $\mathbf{X}$, and $\beta \equiv (\alpha, \beta_1, \ldots, \beta_k)'$ is the parameter vector. The probability of the data conditional on $\mathbf{X}$ is, therefore,

$$p(y_1, \ldots, y_n|\mathbf{X}) = \prod_{i=1}^{n}[\exp(\mathbf{x}_i'\beta)]^{y_i}\left[\frac{1}{1 + \exp(\mathbf{x}_i'\beta)}\right]$$

and the log-likelihood function is

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i\mathbf{x}_i'\beta - \sum_{i=1}^{n}\log_e[1 + \exp(\mathbf{x}_i'\beta)]$$

The partial derivatives of the log likelihood with respect to β are

$$\begin{aligned}\frac{\partial \log_e L(\beta)}{\partial \beta} &= \sum_{i=1}^{n} Y_i\mathbf{x}_i - \sum_{i=1}^{n}\left[\frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}\right]\mathbf{x}_i \\ &= \sum_{i=1}^{n} Y_i\mathbf{x}_i - \sum_{i=1}^{n}\left[\frac{1}{1 + \exp(-\mathbf{x}_i'\beta)}\right]\mathbf{x}_i \qquad (14.15)\end{aligned}$$

Setting the vector of partial derivatives to **0** to maximize the likelihood yields estimating equations

$$\sum_{i=1}^{n}\left[\frac{1}{1+\exp(-\mathbf{x}_i'\mathbf{b})}\right]\mathbf{x}_i = \sum_{i=1}^{n} Y_i\mathbf{x}_i \tag{14.16}$$

where $\mathbf{b} = (A, B_1, \ldots, B_k)'$ is the vector of maximum-likelihood estimates.

The estimating equations (Equation 14.16) have the following intuitive justification:

$$P_i \equiv \frac{1}{1+\exp(-\mathbf{x}_i'\mathbf{b})}$$

is the *fitted* probability for observation i (i.e., the estimated value of π_i). The estimating equations, therefore, set the "fitted sum" $\sum P_i\mathbf{x}_i$ equal to the corresponding observed sum $\sum Y_i\mathbf{x}_i$. In matrix form, we can write the estimating equations as $\mathbf{X}'\mathbf{p} = \mathbf{X}'\mathbf{y}$, where $\mathbf{p} = (P_1, \ldots, P_n)'$ is the vector of fitted values. Note the essential similarity to the least-squares estimating equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$, which can be written $\mathbf{X}'\widehat{\mathbf{y}} = \mathbf{X}'\mathbf{y}$.

Because $\mathbf{b}$ is a maximum-likelihood estimator, its estimated asymptotic covariance matrix can be obtained from the inverse of the information matrix[33]

$$\mathcal{J}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]$$

evaluated at $\boldsymbol{\beta} = \mathbf{b}$. Differentiating Equation 14.15 and making the appropriate substitutions,[34]

$$\begin{aligned}\widehat{\mathcal{V}}(\mathbf{b}) &= \left\{\sum_{i=1}^{n}\frac{\exp(-\mathbf{x}_i'\mathbf{b})}{[1+\exp(-\mathbf{x}_i'\mathbf{b})]^2}\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\\ &= \left[\sum_{i=1}^{n} P_i(1-P_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\\ &= (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\end{aligned}$$

where $\mathbf{V} \equiv \text{diag}\{P_i(1-P_i)\}$ contains the estimated variances of the Y_is. The square roots of the diagonal entries of $\widehat{\mathcal{V}}(\mathbf{b})$ are the asymptotic standard errors, which can be used, as described in the previous section, for inferences about individual parameters of the logit model.

As for the linear model estimated by least squares, general linear hypotheses about the parameters of the logit model can be formulated as $H_0{:}\,\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{L}$ is a $(q \times k+1)$ hypothesis matrix of rank $q \le k+1$ and $\mathbf{c}$ is a $q \times 1$ vector of fixed elements, typically 0.[35] Then the Wald statistic

$$Z_0^2 = (\mathbf{Lb}-\mathbf{c})'\big[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}'\big]^{-1}(\mathbf{Lb}-\mathbf{c})$$

follows an asymptotic chi-square distribution with q degrees of freedom under the hypothesis H_0. For example, to test the omnibus hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$, we take

$$\underset{(k\times k+1)}{\mathbf{L}} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix} = [\underset{(k\times 1)}{\mathbf{0}}, \mathbf{I}_k]$$

and $\mathbf{c} = \underset{(k\times 1)}{\mathbf{0}}$.

[33] See Appendix D on probability and estimation.

[34] See Exercise 14.8.

[35] See Section 9.4.3.

Likewise, the asymptotic $100(1-a)\%$ joint confidence region for a subset of q parameters $\boldsymbol{\beta}_1$ takes the form

$$(\mathbf{b}_1 - \boldsymbol{\beta}_1)'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \boldsymbol{\beta}_1) \leq \chi^2_{q,a}$$

Here, $\mathbf{V}_{11}$ is the $(q \times q)$ submatrix of $\widehat{\mathcal{V}}(\mathbf{b})$ that pertains to the estimates $\mathbf{b}_1$, and $\chi^2_{q,a}$ is the critical value of the chi-square distribution for q degrees of freedom with probability a to the right.

Unlike the normal equations for a linear model, the logit-model estimating equations (Equation 14.16) are nonlinear functions of $\mathbf{b}$ and, therefore, require iterative solution. One common approach to solving the estimating equations is the *Newton-Raphson method*, which can be described as follows:[36]

1. Select initial estimates $\mathbf{b}_0$; a simple choice is $\mathbf{b}_0 = \mathbf{0}$.
2. At each iteration $l+1$, compute new estimates

$$\mathbf{b}_{l+1} = \mathbf{b}_l + (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p}_l) \tag{14.17}$$

where $\mathbf{p}_l \equiv \{1/[1+\exp(-\mathbf{x}_i'\mathbf{b}_l)]\}$ is the vector of fitted values from the previous iteration and $\mathbf{V}_l \equiv \text{diag}\{P_{li}(1-P_{li})\}$.

3. Iterations continue until $\mathbf{b}_{l+1} \approx \mathbf{b}_l$ to the desired degree of accuracy. When convergence takes place,

$$(\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p}_l) \approx \mathbf{0}$$

and thus the estimating equations $\mathbf{X}'\mathbf{p} = \mathbf{X}'\mathbf{y}$ are approximately satisfied. Conversely, if the fitted sums $\mathbf{X}'\mathbf{p}_l$ are very different from the observed sums $\mathbf{X}'\mathbf{y}$, then there will be a large adjustment in $\mathbf{b}$ at the next iteration. The Newton-Raphson procedure conveniently produces the estimated asymptotic covariance matrix of the coefficients $\widehat{\mathcal{V}}(\mathbf{b}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$ as a by-product.

Suppose, now, that we have obtained complete convergence of the Newton-Raphson procedure to the maximum-likelihood estimator $\mathbf{b}$. From Equation 14.17, we have

$$\mathbf{b} = \mathbf{b} + (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p})$$

which we can rewrite as

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}^*$$

where[37]

$$\mathbf{y}^* \equiv \mathbf{X}\mathbf{b} + \mathbf{V}^{-1}(\mathbf{y} - \mathbf{p})$$

These formulas suggest an analogy between maximum-likelihood estimation of the linear logit model and weighted-least-squares regression. The analogy is the basis of an alternative method for calculating the maximum-likelihood estimates called *iterative weighted least squares* (IWLS):[38]

[36] This approach was first applied by R. A. Fisher, in the context of a probit model, and is sometimes termed *Fisher scoring* in his honor.

[37] See Exercise 14.9.

[38] This method is also called *iteratively reweighted least squares* (IRLS). See Section 12.2.2 for an explanation of weighted-least-squares estimation. In fact, the IWLS algorithm is an alternative implementation of the Newton-Raphson method and leads to the same history of iterations.

1. As before, select arbitrary initial values $\mathbf{b}_0$.
2. At each iteration l, calculate fitted values $\mathbf{p}_l \equiv \{1/[1 + \exp(-\mathbf{x}_i'\mathbf{b}_l)]\}$, the variance matrix $\mathbf{V}_l \equiv \text{diag}\{P_{li}(1 - P_{li})\}$, and the "pseudoresponse variable" $\mathbf{y}_l^* \equiv \mathbf{X}\mathbf{b}_l + \mathbf{V}_l^{-1}(\mathbf{y} - \mathbf{p}_l)$.
3. Calculate updated estimates by weighted-least-squares regression of the pseudoresponse on the Xs, using the current variance matrix for weights:

$$\mathbf{b}_{l+1} = (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_l\mathbf{y}_l^*$$

4. Repeat Steps 2 and 3 until the coefficients converge.

14.2 Models for Polytomous Data

A limitation of the logit and probit models of the previous section is that they apply only to dichotomous response variables. In the Chilean plebiscite data, for example, many of the voters surveyed indicated that they were undecided, and some said that they planned to abstain or refused to reveal their voting intentions. Polytomous data of this sort are common, and it is desirable to model them in a natural manner—not simply to ignore some of the categories (e.g., restricting attention to those who responded *yes* or *no*) or to combine categories arbitrarily to produce a dichotomy.

In this section, I will describe three general approaches to modeling polytomous data:[39]

1. modeling the polytomy directly as a set of unordered categories, using a generalization of the dichotomous logit model;
2. constructing a set of nested dichotomies from the polytomy, fitting an independent logit or probit model to each dichotomy; and
3. extending the unobserved-variable interpretation of the dichotomous logit and probit models to ordered polytomies.

14.2.1 The Polytomous Logit Model

It is possible to generalize the dichotomous logit model to a polytomy by employing the multivariate logistic distribution. This approach has the advantage of treating the categories of the polytomy in a nonarbitrary, symmetric manner (but the disadvantage that the analysis is relatively complex).[40]

Suppose that the response variable Y can take on any of m qualitative values, which, for convenience, we number $1, 2, \ldots, m$. To anticipate the example employed in this section, a voter in the 2001 British election voted for (1) the Labour Party, (2) the Conservative Party, or (3) the Liberal Democrats. Although the categories of Y are numbered, we do not, in general, attribute ordinal properties to these numbers: They are simply category *labels*. Let π_{ij} denote the probability that the ith observation falls in the jth category of the response variable; that is, $\pi_{ij} \equiv \Pr(Y_i = j)$, for $j = 1, \ldots, m$.

[39] Additional statistical models for polytomous data are described, for example, in Agresti (2002).

[40] A similar probit model based on the multivariate-normal distribution is somewhat more difficult to estimate because of the necessity of evaluating a multivariate integral, but is sometimes preferred to the polytomous logit model developed in this section (see Exercise 14.12).

We have available k regressors, $X_1, \ldots, X_k$, on which the π_{ij} depend. More specifically, suppose that this dependence can be modeled using the *multivariate logistic distribution*:

$$\pi_{ij} = \frac{\exp(\gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik})}{1 + \sum_{l=1}^{m-1} \exp(\gamma_{0l} + \gamma_{1l}X_{i1} + \cdots + \gamma_{kl}X_{ik})} \quad \text{for } j = 1, \ldots, m-1 \tag{14.18}$$

$$\pi_{im} = 1 - \sum_{l=1}^{m-1} \pi_{ij} \quad \text{(for category } m\text{)}$$

This model is sometimes called the *multinomial logit model.*[41] There is, then, one set of parameters, $\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{kj}$, for each response category but the last. The last category (i.e., category m) functions as a type of baseline. The use of a baseline category is one way of avoiding redundant parameters because of the restriction, reflected in the second part of Equation 14.18, that the response category probabilities for each observation must sum to 1:[42]

$$\sum_{j=1}^{m} \pi_{ij} = 1$$

The denominator of π_{ij} in the first line of Equation 14.18 imposes this restriction.

Some algebraic manipulation of Equation 14.18 produces[43]

$$\log_e \frac{\pi_{ij}}{\pi_{im}} = \gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik} \quad \text{for } j = 1, \ldots, m-1$$

The regression coefficients, therefore, represent effects on the log-odds of membership in category j versus the baseline category. It is also possible to form the log-odds of membership in *any* pair of categories j and j' (other than category m):

$$\begin{aligned} \log_e \frac{\pi_{ij}}{\pi_{ij'}} &= \log_e \left(\frac{\pi_{ij}/\pi_{im}}{\pi_{ij'}/\pi_{im}} \right) \\ &= \log_e \frac{\pi_{ij}}{\pi_{im}} - \log_e \frac{\pi_{ij'}}{\pi_{im}} \\ &= (\gamma_{0j} - \gamma_{0j'}) + (\gamma_{1j} - \gamma_{1j'})X_{i1} + \cdots + (\gamma_{kj} - \gamma_{kj'})X_{ik} \end{aligned} \tag{14.19}$$

Thus, the regression coefficients for the logit between any pair of categories are the *differences* between corresponding coefficients for the two categories.

[41] I prefer to reserve the term *multinomial logit model* for a version of the model that can accommodate counts for the several categories of the response variable in a contingency table formed by discrete explanatory variables. I make a similar distinction between *binary* and *binomial* logit models, with the former term applied to individual observations and the latter to counts of "successes" and "failures" for a dichotomous response. See the discussion of the application of logit models to contingency tables in Section 14.3.

[42] An alternative is to treat the categories symmetrically:

$$\pi_{ij} = \frac{\exp(\gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik})}{\sum_{l=1}^{m} \exp(\gamma_{0l} + \gamma_{1l}X_{i1} + \cdots + \gamma_{kl}X_{ik})}$$

but to impose a linear restriction—analogous to a sigma constraint in an ANOVA model (see Chapter 8)—on the parameters of the model. This approach produces somewhat more difficult computations, however, and has no real advantages. Although the choice of baseline category is essentially arbitrary and inconsequential, if one of the response categories represents a natural point of comparison, one might as well use it as the baseline.

[43] See Exercise 14.10.

To gain further insight into the polytomous logit model, suppose that the model is specialized to a dichotomous response variable. Then, $m = 2$, and

$$\log_e \frac{\pi_{i1}}{\pi_{i2}} = \log_e \frac{\pi_{i1}}{1 - \pi_{i1}} = \gamma_{01} + \gamma_{11} X_{i1} + \cdots + \gamma_{k1} X_{ik}$$

When it is applied to a dichotomy, the polytomous logit model is, therefore, identical to the dichotomous logit model of the previous section.

The following example is adapted from work by Andersen, Heath, and Sinnott (2002) on the 2001 British election, using data from the final wave of the 1997–2001 British Election Panel Study (BEPS) (also see Fox & Andersen, 2006). The central issue addressed in the data analysis is the potential interaction between respondents' political knowledge and political attitudes in determining their vote. The response variable, vote, has three categories: Labour, Conservative, and Liberal Democrat; individuals who voted for smaller parties are excluded from the analysis. There are several explanatory variables:

- Attitude toward European integration, an 11-point scale, with high scores representing a negative attitude (so-called Euro-scepticism).
- Knowledge of the platforms of the three parties on the issue of European integration, with integer scores ranging from 0 through 3. (Labour and the Liberal Democrats supported European integration, the Conservatives were opposed.)
- Other variables included in the model primarily as "controls"—age, gender, perceptions of national and household economic conditions, and ratings of the three party leaders.

The coefficients of a polytomous logit model fit to the BEPS data are shown, along with their standard errors, in Table 14.5. This model differs from those I have described previously in this text in that it includes the product of two quantitative explanatory variables, representing the *linear-by-linear interaction* between these variables:[44] Focusing on the Conservative/Liberal-Democrat logit, for example, when political knowledge is 0, the slope for attitude toward European integration ("Europe") is −0.068. With each unit increase in political knowledge, the slope for Europe increases by 0.183, thus becoming increasingly positive. This result is sensible: Those with more knowledge of the parties' positions are more likely to vote in conformity with their own position on the issue. By the same token, at low levels of Europe, the slope for political knowledge is negative, but it increases by 0.183 with each unit increase in Europe. By a Wald test, this interaction coefficient is highly statistically significant, with $Z = 0.183/0.028 = 6.53$, for which $p \ll .0001$.

An analysis-of-deviance table for the model appears in Table 14.6. Note that each term has two degrees of freedom, representing the two coefficients for the term, one for the Labour/Liberal-Democrat logit and the other for the Conservative/Liberal-Democrat logit. All the terms in the model are highly statistically significant, with the exception of gender and perception of household economic position.

Although we can therefore try to understand the fitted model by examining its coefficients, there are two obstacles to doing so: (1) As explained, the interaction between political knowledge and attitude toward European integration requires that we perform mental gymnastics to combine the estimated coefficient for the interaction with the coefficients for the "main-effect" regressors that are marginal to the interaction. (2) The structure of the polytomous logit model, which is for log-odds of pairs of categories (each category versus the baseline Liberal-Democrat category), makes it difficult to formulate a general understanding of the results.

[44] For more on models of this form, see Section 17.1 on polynomial regression.

Table 14.5 Polytomous Logit Model Fit to the BEPS Data

	Labour/Liberal Democrat	
Coefficient	*Estimate*	*Standard Error*
Constant	−0.155	0.612
Age	−0.005	0.005
Gender (male)	0.021	0.144
Perception of Economy	0.377	0.091
Perception of Household Economic Position	0.171	0.082
Evaluation of Blair (Labour leader)	0.546	0.071
Evaluation of Hague (Conservative leader)	−0.088	0.064
Evaluation of Kennedy (Liberal Democrat leader)	−0.416	0.072
Attitude Toward European Integration	−0.070	0.040
Political Knowledge	−0.502	0.155
Europe × Knowledge	0.024	0.021

	Conservative/Liberal Democrat	
Coefficient	*Estimate*	*Standard Error*
Constant	0.718	0.734
Age	0.015	0.006
Gender (male)	−0.091	0.178
Perception of Economy	−0.145	0.110
Perception of Household Economic Position	−0.008	0.101
Evaluation of Blair (Labour leader)	−0.278	0.079
Evaluation of Hague (Conservative leader)	0.781	0.079
Evaluation of Kennedy (Liberal Democrat leader)	−0.656	0.086
Attitude Toward European Integration	−0.068	0.049
Political Knowledge	−1.160	0.219
Europe × Knowledge	0.183	0.028

Table 14.6 Analysis of Deviance for the Polytomous Logit Model Fit to the BEPS Data

Source	*df*	G_0^2	*p*
Age	2	13.87	.0009
Gender	2	0.45	.78
Perception of Economy	2	30.60	$\ll$.0001
Perception of Household Economic Position	2	5.65	.059
Evaluation of Blair	2	135.37	$\ll$.0001
Evaluation of Hague	2	166.77	$\ll$.0001
Evaluation of Kennedy	2	68.88	$\ll$.0001
Attitude Toward European Integration	2	78.03	$\ll$.0001
Political Knowledge	2	55.57	$\ll$.0001
Europe × Knowledge	2	50.80	$\ll$.0001

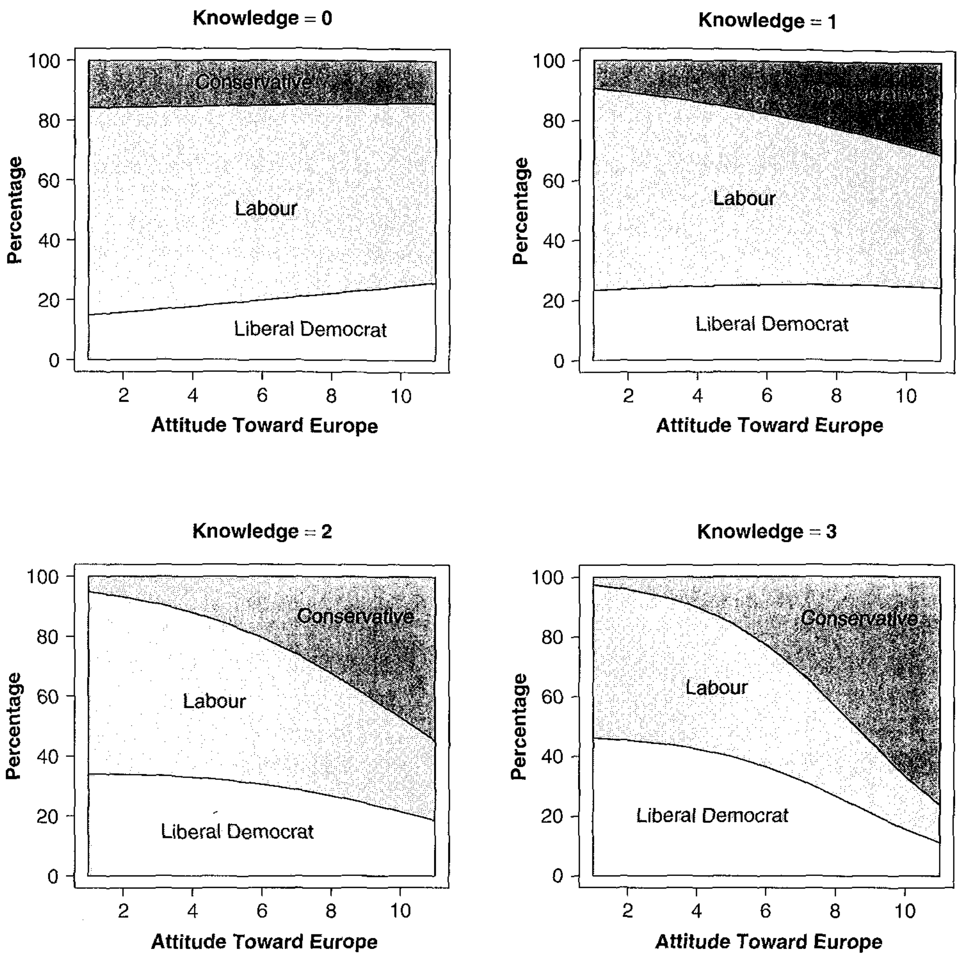

Figure 14.6 Effect display for the interaction between attitude toward European integration and political knowledge in the polytomous logit model for vote in the 2001 British election.

Once more, a graphical representation of the fitted model can greatly aid in its interpretation. An effect plot for the interaction of attitude toward European integration with political knowledge is shown in Figure 14.6. The strategy for constructing this plot is the usual one, adapted to the polytomous logit model: Compute the fitted probability of membership in each of the three categories of the response variable, letting Europe and knowledge range in combination over their values, while the other explanatory variables are fixed to average values. It is apparent that as political knowledge increases, vote conforms more closely to the respondent's attitude toward European integration.

*Details of Estimation**

To fit the polytomous logit model given in Equation 14.18 (on page 356) to data, we may again invoke the method of maximum likelihood. Recall that each Y_i takes on its possible values $1, 2, \ldots, m$ with probabilities $\pi_{i1}, \pi_{i2}, \ldots, \pi_{im}$. Following Nerlove and Press (1973), let us define indicator variables $W_{i1}, \ldots, W_{im}$ so that $W_{ij} = 1$ if $Y_i = j$, and $W_{ij} = 0$ if $Y_i \neq j$; thus,

$$\begin{aligned} p(y_i) &= \pi_{i1}^{w_{i1}}\,\pi_{i2}^{w_{i2}}\cdots\pi_{im}^{w_{im}} \\ &= \prod_{j=1}^{m}\pi_{ij}^{w_{ij}} \end{aligned}$$

If the observations are sampled independently, then their joint probability distribution is given by

$$\begin{aligned} p(y_1,\ldots,y_n) &= p(y_1)\times\cdots\times p(y_n) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m}\pi_{ij}^{w_{ij}} \end{aligned}$$

For compactness, define the following vectors:

$$\begin{aligned} \mathbf{x}_i' &\equiv (1, X_{i1},\ldots,X_{ik}) \\ \boldsymbol{\gamma}_j &\equiv (\gamma_{0j},\gamma_{1j},\ldots,\gamma_{kj})' \end{aligned}$$

and the model matrix

$$\underset{(n\times k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

It is convenient to impose the restriction $\sum_{j=1}^{m}\pi_{ij} = 1$ by setting $\boldsymbol{\gamma}_m = \mathbf{0}$ (making category m the baseline, as explained previously). Then, employing Equation 14.18,

$$p(y_1,\ldots,y_n|\mathbf{X}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[\frac{\exp(\mathbf{x}_i'\boldsymbol{\gamma}_j)}{1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)}\right]^{w_{ij}} \tag{14.20}$$

and the log likelihood is

$$\begin{aligned} \log_e L(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{m-1}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\left\{\mathbf{x}_i'\boldsymbol{\gamma}_j - \log_e\left[1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)\right]\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m-1} W_{ij}\mathbf{x}_i'\boldsymbol{\gamma}_j - \sum_{i=1}^{n}\log_e\left[1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)\right] \end{aligned}$$

because $\sum_{j=1}^{m} W_{ij} = 1$ and $\boldsymbol{\gamma}_m = \mathbf{0}$; setting $\boldsymbol{\gamma}_m = \mathbf{0}$ accounts for the 1 in the denominator of Equation 14.20 because $\exp(\mathbf{x}_i'\mathbf{0}) = 1$.

Differentiating the log likelihood with respect to the parameters, and setting the partial derivatives to $\mathbf{0}$, produces the nonlinear estimating equations:[45]

$$\begin{aligned} \sum_{i=1}^{n} W_{ij}\mathbf{x}_i &= \sum_{i=1}^{n}\mathbf{x}_i\frac{\exp(\mathbf{x}_i'\mathbf{c}_j)}{1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\mathbf{c}_l)} \quad \text{for } j=1,\ldots,m-1 \\ &= \sum_{i=1}^{n} P_{ij}\mathbf{x}_i \end{aligned} \tag{14.21}$$

[45] See Exercise 14.11.

where $\mathbf{c}_j \equiv \widehat{\boldsymbol{\gamma}}_j$ are the maximum-likelihood estimators of the regression coefficients, and the

$$P_{ij} \equiv \frac{\exp(\mathbf{x}_i'\mathbf{c}_j)}{1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\mathbf{c}_l)}$$

are the fitted probabilities. As in the dichotomous logit model, the maximum-likelihood estimator sets observed sums equal to fitted sums. The estimating equations (Equation 14.21) are nonlinear and, therefore, require iterative solution.

Let us stack up all the parameters in a large vector:

$$\underset{[(m-1)(k+1)\times 1]}{\boldsymbol{\gamma}} \equiv \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_{m-1} \end{bmatrix}$$

The information matrix is[46]

$$\underset{[(m-1)(k+1)\times(m-1)(k+1)]}{\mathcal{J}(\boldsymbol{\gamma})} = \begin{bmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} & \cdots & \mathcal{J}_{1,m-1} \\ \mathcal{J}_{21} & \mathcal{J}_{22} & \cdots & \mathcal{J}_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{J}_{m-1,1} & \mathcal{J}_{m-1,2} & \cdots & \mathcal{J}_{m-1,m-1} \end{bmatrix}$$

where

$$\begin{aligned}\underset{[(k+1)\times(k+1)]}{\mathcal{J}_{jj}} &= -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}_j'}\right] \\ &= \sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'\exp(\mathbf{x}_i'\boldsymbol{\gamma}_j)[1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)-\exp(\mathbf{x}_i'\boldsymbol{\gamma}_j)]}{[1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)]^2}\end{aligned} \tag{14.22}$$

and

$$\begin{aligned}\underset{[(k+1)\times(k+1)]}{\mathcal{J}_{jj'}} &= -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}_j\,\partial\boldsymbol{\gamma}_{j'}'}\right] \\ &= -\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'\exp[\mathbf{x}_i'(\boldsymbol{\gamma}_{j'}+\boldsymbol{\gamma}_j)]}{[1+\sum_{l=1}^{m-1}\exp(\mathbf{x}_i'\boldsymbol{\gamma}_l)]^2}\end{aligned} \tag{14.23}$$

The estimated asymptotic covariance matrix of

$$\mathbf{c} \equiv \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{m-1} \end{bmatrix}$$

is obtained from the inverse of the information matrix, replacing $\boldsymbol{\gamma}$ with $\mathbf{c}$.

14.2.2 Nested Dichotomies

Perhaps the simplest approach to polytomous data—because it employs the already-familiar dichotomous logit or probit model—is to fit separate models to each of a set of dichotomies derived

[46] See Exercise 14.11.

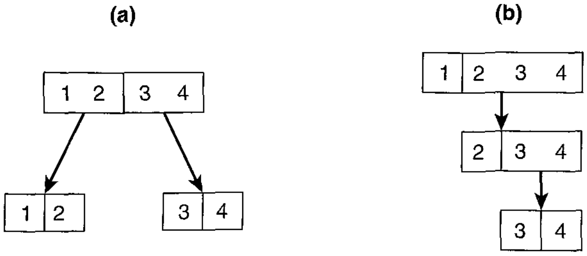

Figure 14.7 Alternative sets of nested dichotomies [(a) and (b)] for a four-category polytomous response variable.

from the polytomy. These dichotomies are constructed so that the likelihood for the polytomous response variable is the product of the likelihoods for the dichotomies—that is, the models are statistically independent even though they are fitted to data from the same sample. The likelihood is separable in this manner if the set of dichotomies is *nested*.[47] Although the system of nested dichotomies constitutes a model for the polytomy, and although this model often yields fitted probabilities that are very similar to those associated with the polytomous logit model of the previous section, the two models are not equivalent.

A nested set of $m - 1$ dichotomies is produced from an m-category polytomy by successive binary partitions of the categories of the polytomy. Two examples for a four-category variable are shown in Figure 14.7. In part (a) of this figure, the dichotomies are $\{12, 34\}$ (i.e., the combination of Categories 1 and 2 vs. the combination of Categories 3 and 4); $\{1, 2\}$ (Category 1 vs. Category 2); and $\{3, 4\}$ (Category 3 vs. Category 4). In part (b), the nested dichotomies are $\{1, 234\}$, $\{2, 34\}$, and $\{3, 4\}$. This simple—and abstract—example illustrates a key property of nested dichotomies: The set of nested dichotomies selected to represent a polytomy is *not* unique. Because the results of the analysis and their interpretation depend on the set of nested dichotomies that is selected, this approach to polytomous data is reasonable only when a particular choice of dichotomies is substantively compelling. If the dichotomies are purely arbitrary, or if alternative sets of dichotomies are equally reasonable and interesting, then nested dichotomies should probably not be used to analyze the data.

Nested dichotomies are an especially attractive approach when the categories of the polytomy represent ordered progress through the stages of a process. Imagine, for example, that the categories in Figure 14.7(b) represent adults' attained level of education: (1) less than high school; (2) high-school graduate; (3) some postsecondary; (4) postsecondary degree. Because individuals normally progress through these categories in sequence, the dichotomy $\{1, 234\}$ represents the completion of high school; $\{2, 34\}$ the continuation to postsecondary education, conditional on high school graduation; and $\{3, 4\}$ the completion of a degree conditional on undertaking a postsecondary education.[48]

*Why Nested Dichotomies Are Independent**

For simplicity, I will demonstrate the independence of the nested dichotomies $\{12, 3\}$ and $\{1, 2\}$. By repeated application, this result applies generally to any system of nested dichotomies. Let W_{i1}, W_{i2}, and W_{i3} be dummy variables indicating whether the polytomous response variable

[47] A proof of this property of nested dichotomies will be given presently.

[48] Fienberg (1980, pp. 110–116) terms ratios of odds formed from these nested dichotomies *continuation ratios*. An example employing nested dichotomies for educational attainment is developed in the data-analysis exercises for this chapter.

Y_i is 1, 2, or 3. For example, $W_{i1} = 1$ if $Y_i = 1$, and 0 otherwise. Let Y_i' be a dummy variable representing the first dichotomy, $\{12, 3\}$: That is, $Y_i' = 1$ when $Y_i = 1$ or 2, and $Y_i' = 0$ when $Y_i = 3$. Likewise, let Y_i'' be a dummy variable representing the second dichotomy, $\{1, 2\}$: $Y_i'' = 1$ when $Y_i = 1$, and $Y_i'' = 0$ when $Y_i = 2$; Y_i'' is *undefined* when $Y_i = 3$. We need to show that $p(y_i) = p(y_i')p(y_i'')$. [To form this product, we adopt the convention that $p(y_i'') \equiv 1$ when $Y_i = 3$.]

The probability distribution of Y_i' is given by

$$\begin{aligned} p(y_i') &= (\pi_{i1} + \pi_{i2})^{y_i'} \pi_{i3}^{1-y_i'} \\ &= (\pi_{i1} + \pi_{i2})^{w_{i1}+w_{i2}} \pi_{i3}^{w_{i3}} \end{aligned} \tag{14.24}$$

where $\pi_{ij} \equiv \Pr(Y_i = j)$ for $j = 1, 2, 3$. To derive the probability distribution of Y_i'', note that

$$\begin{aligned} \Pr(Y_i'' = 1) &= \Pr(Y_i = 1 | Y_i \neq 3) = \frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}} \\ \Pr(Y_i'' = 0) &= \Pr(Y_i = 2 | Y_i \neq 3) = \frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}} \end{aligned}$$

and, thus,

$$\begin{aligned} p(y_i'') &= \left(\frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}}\right)^{y_i''} \left(\frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}}\right)^{1-y_i''} \\ &= \left(\frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}}\right)^{w_{i1}} \left(\frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}}\right)^{w_{i2}} \end{aligned} \tag{14.25}$$

Multiplying Equation 14.24 by Equation 14.25 produces

$$p(y_i')p(y_i'') = \pi_{i1}^{w_{i1}} \pi_{i2}^{w_{i2}} \pi_{i3}^{w_{i3}} = p(y_i)$$

which is the required result.

Because the dichotomies Y' and Y'' are independent, it is legitimate to combine models for these dichotomies to form a model for the polytomy Y. Likewise, we can sum likelihood-ratio or Wald test statistics for the two dichotomies.

14.2.3 Ordered Logit and Probit Models

Imagine (as in Section 14.1.3) that there is a latent (i.e., unobservable) variable ξ that is a linear function of the Xs plus a random error:

$$\xi_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Now, however, suppose that instead of dividing ξ into two regions to produce a dichotomous response, ξ is dissected by $m - 1$ *thresholds* (i.e., boundaries) into m regions. Denoting the thresholds by $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$, and the resulting response by Y, we observe

$$Y_i = \begin{cases} 1 & \text{if } \xi_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < \xi_i \leq \alpha_2 \\ \vdots & \\ m-1 & \text{if } \alpha_{m-2} < \xi_i \leq \alpha_{m-1} \\ m & \text{if } \alpha_{m-1} < \xi_i \end{cases} \tag{14.26}$$

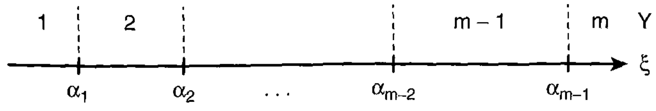

Figure 14.8 The thresholds $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$ divide the latent continuum ξ into m regions, corresponding to the values of the observable variable Y.

The thresholds, regions, and corresponding values of ξ and Y are represented graphically in Figure 14.8. Note that the thresholds are not in general uniformly spaced.

Using Equation 14.26, we can determine the cumulative probability distribution of Y:

$$\begin{aligned}\Pr(Y_i \le j) &= \Pr(\xi_i \le \alpha_j)\\ &= \Pr(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \le \alpha_j)\\ &= \Pr(\varepsilon_i \le \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})\end{aligned}$$

If the errors ε_i are independently distributed according to the standard normal distribution, then we obtain the ordered probit model.[49] If the errors follow the similar logistic distribution, then we get the ordered logit model. In the latter event,

$$\begin{aligned}\text{logit}\,[\Pr(Y_i \le j)] &= \log_e \frac{\Pr(Y_i \le j)}{\Pr(Y_i > j)}\\ &= \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\end{aligned}$$

Equivalently,

$$\begin{aligned}\text{logit}\,[\Pr(Y_i > j)] &= \log_e \frac{\Pr(Y_i > j)}{\Pr(Y_i \le j)}\\ &= (\alpha - \alpha_j) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}\end{aligned} \tag{14.27}$$

for $j = 1, 2, \ldots, m-1$.

The logits in Equation 14.27 are for cumulative categories—at each point contrasting categories above category j with category j and below. The slopes for each of these regression equations are identical; the equations differ only in their intercepts. The logistic-regression surfaces are, therefore, horizontally parallel to each other, as illustrated in Figure 14.9 for $m = 4$ response categories and a single X. (For the more general case, just replace X by the linear predictor $\eta = \beta_1 X_1 + \cdots + \beta_k X_k$.)

Put another way, for a fixed set of Xs, any two different cumulative log-odds (i.e., logits)—say, at categories j and j'—differ only by the constant $(\alpha_{j'} - \alpha_j)$. The odds, therefore, are *proportional* to one another; that is,

$$\frac{\text{odds}_j}{\text{odds}_{j'}} = \exp\left(\text{logit}_j - \text{logit}_{j'}\right) = \exp(\alpha_{j'} - \alpha_j) = \frac{e^{\alpha_{j'}}}{e^{\alpha_j}}$$

where, for example, $\text{odds}_j \equiv \Pr(Y_i > j)$ and $\text{logit}_j \equiv \text{logit}\,[\Pr(Y_i > j)]$. For this reason, Equation 14.27 is called the *proportional-odds logit model*.

There are $(k+1) + (m-1) = k + m$ parameters to estimate in the proportional-odds model, including the regression coefficients α, $\beta_1, \ldots, \beta_k$ and the category thresholds $\alpha_1, \ldots, \alpha_{m-1}$.

[49] As in the dichotomous case, we conveniently fix the error variance to 1 to set the scale of the latent variable ξ. The resulting ordered probit model does not have the proportional-odds property described below.

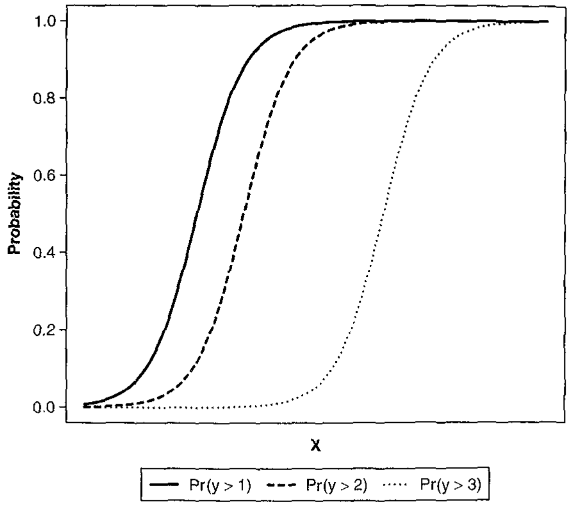

Figure 14.9 The proportional-odds model for four response categories and a single explanatory variable X. The logistic regression curves are horizontally parallel.

SOURCE: Adapted from Agresti (1990, fig. 9.1), *Categorical Data Analysis*. Copyright ©1990 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

Note, however, that there is an extra parameter in the regression equations (Equation 14.27) because each equation has its own constant, $-\alpha_j$, along with the common constant α. A simple solution is to set $\alpha = 0$ (and to absorb the negative sign into α_j), producing[50]

$$\text{logit}\,[\Pr(Y_i > j)] = \alpha_j + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} \qquad (14.28)$$

In this parametrization, the intercepts α_j are the *negatives* of the category thresholds.

Figure 14.10 illustrates the proportional-odds model for $m = 4$ response categories and a single X. The conditional distribution of the latent variable ξ is shown for two representative values of the explanatory variable, x_1 [where $\Pr(Y > 3) = \Pr(Y = 4)$ is about .2] and x_2 [where $\Pr(Y = 4)$ is about .98]. McCullagh (1980) explains how Equation 14.27 can be fit by the method of maximum likelihood (and discusses alternatives to the proportional-odds model).

To illustrate the use of the proportional-odds model, I draw on data from the World Values Survey (WVS) of 1995–1997 (European Values Study Group and World Values Survey Association, 2000).[51] Although the WVS collects data in many countries, to provide a manageable example, I will restrict attention to only four: Australia, Sweden, Norway, and the United States. The combined sample size for these four countries is 5,381. The response variable in the analysis is the answer to the question "Do you think that what the government is doing for people in poverty is about the right amount, too much, or too little?" There are, therefore, three ordered categories: *too little*, *about right*, *too much*. There are several explanatory variables: gender (represented by a dummy variable coded 1 for *men* and 0 for *women*); whether or not the respondent belonged to a religion (coded 1 for *yes*, 0 for *no*); whether or not the respondent had a university degree (coded 1 for *yes* and 0 for *no*); age (in years, ranging from 18 to 87); and country (entered into the

[50] Setting $\alpha = 0$ implicitly establishes the origin of the latent variable ξ (just as fixing the error variance establishes its unit of measurement). An alternative would be to fix one of the thresholds to 0. These choices are arbitrary and inconsequential.

[51] This illustration is adapted from Fox and Andersen (2006).

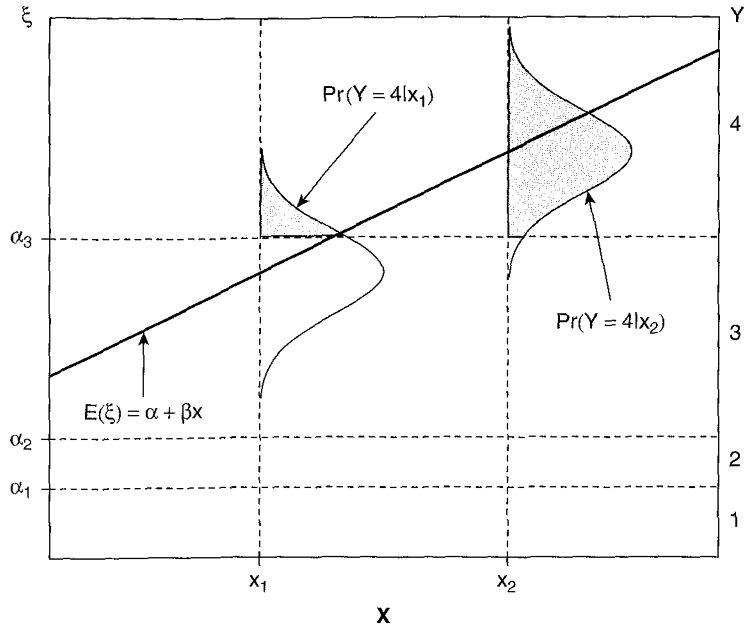

Figure 14.10 The proportional-odds model for four response categories and a single explanatory variable X. The latent response variable ξ has a linear regression on X. The latent continuum ξ and thresholds α_j appear at the left of the graph, the observable response Y at the right. The conditional logistic distribution of the latent variable is shown for two values of the explanatory variable, x_1 and x_2. The shaded area in each distribution gives the conditional probability that $Y = 4$.

SOURCE: Adapted from Agresti (1990, fig. 9.2), *Categorical Data Analysis*. Copyright ©1990 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

model as a set of three dummy regressors, with *Australia* as the baseline category). Preliminary analysis of the data suggested a roughly linear age effect.

Table 14.7 shows the analysis of deviance for an initial model fit to the data incorporating interactions between country and each of the other explanatory variables. As usual, the likelihood-ratio tests in the table are computed by contrasting the deviances for alternative models, with and without the terms in question. These tests were formulated in conformity with the principle of marginality. So, for example, the test for the country-by-age interaction was computed by dropping this term from the full model, and the test for the country main effect was computed by dropping the dummy regressors for country from a model that includes only main effects.

With the exception of the interaction between country and gender, all these interactions prove to be statistically significant. Estimated coefficients and their standard errors for a final model, removing the nonsignificant interaction between country and gender, appear in Table 14.8. This table also shows the estimated thresholds between response categories, which are, as explained, the negatives of the intercepts of the proportional-odds model.

Interpretation of the estimated coefficients for the proportional-odds model in Table 14.8 is complicated by the interactions in the model and by the multiple-category response. I will use the interaction between age and country to illustrate: We can see that the age slope is positive in the baseline country of Australia (suggesting that sympathy for the poor declines with age in Australia) and that this slope is nearly zero in Norway, smaller in Sweden than in Australia, and

Table 14.7 Analysis of Deviance Table for the Proportional-Odds Model Fit to the World Values Survey Data

Source	*df*	G_0^2	*p*
Country	3	250.881	$\ll .0001$
Gender	1	10.749	.0010
Religion	1	4.132	.042
Education	1	4.284	.038
Age	1	49.950	$\ll .0001$
Country × Gender	3	3.049	.38
Country × Religion	3	21.143	$<.0001$
Country × Education	3	12.861	.0049
Country × Age	3	17.529	.0005

Table 14.8 Estimated Proportional-Odds Model Fit to the World Values Survey Data

Coefficient	*Estimate*	*Standard Error*
Gender (Men)	0.1744	0.0532
Country (Norway)	0.1516	0.3355
Country (Sweden)	−1.2237	0.5821
Country (United States)	1.2225	0.3068
Religion (Yes)	0.0255	0.1120
Education (Degree)	−0.1282	0.1676
Age	0.0153	0.0026
Country (Norway) × Religion	−0.2456	0.2153
Country (Sweden) × Religion	−0.9031	0.5125
Country (United States) × Religion	0.5706	0.1733
Country (Norway) × Education	0.0524	0.2080
Country (Sweden) × Education	0.6359	0.2141
Country (United States) × Education	0.3103	0.2063
Country (Norway) × Age	−0.0156	0.0044
Country (Sweden) × Age	−0.0090	0.0047
Country (United States) × Age	0.0008	0.0040
Thresholds		
$-\widehat{\alpha}_1$ (Too Little| About Right)	0.7699	0.1491
$-\widehat{\alpha}_2$ (About Right | Too Much)	2.5372	0.1537

very slightly larger in the United States than in Australia, but a more detailed understanding of the age-by-country interaction is hard to discern from the coefficients alone. Figures 14.11 and 14.12 show alternative effect displays of the age-by-country interaction. The strategy for constructing these displays is the usual one—compute fitted values under the model letting age and country range over their values while other explanatory variables (i.e., gender, religion, and education) are held to average values. Figure 14.11 plots the fitted probabilities of response (as percentages) by age for each country; Figure 14.12 plots the fitted value of the latent response variable by age for each country, and shows the intercategory thresholds.

The proportional-odds model of Equation 14.28 (on page 365) constrains corresponding slopes for the $m - 1$ cumulative logits to be equal. By relaxing this strong constraint, and fitting a model to the cumulative logits that permits different slopes along with different intercepts, we can test the proportional-odds assumption:

$$\text{logit}\,[\Pr(Y_i > j)] = \alpha_j + \beta_{j1}X_{i1} + \cdots + \beta_{jk}X_{ik}, \text{ for } j = 1, \ldots, m-1 \tag{14.29}$$

Like the polytomous logit model of Equation 14.18 (on page 356), this new model has $(m - 1)$ $(k + 1)$ parameters, but the two models are for *different* sets of logits. The deviances and numbers of parameters for the three models fit to the World Values Survey data are as follows:

Model	*Residual Deviance*	*Number of Parameters*
Proportional-Odds Model (Equation 14.28)	10,350.12	18
Cumulative Logits, Unconstrained Slopes (Equation 14.29)	9,961.63	34
Polytomous Logit Model (Equation 14.18)	9,961.26	34

The likelihood-ratio statistic for testing the assumption of proportional odds is therefore $G_0^2 = 10{,}350.12 - 9{,}961.63 = 388.49$, on $34 - 18 = 16$ degrees of freedom. This test statistic is highly statistically significant, leading us to reject the proportional-odds assumption for these data. Note that the deviance for the model that relaxes the proportional-odds assumption is nearly identical to the deviance for the polytomous logit model. This is typically the case, in my experience.[52]

14.2.4 Comparison of the Three Approaches

Several approaches can be taken to modeling polytomous data, including (1) modeling the polytomy directly using a logit model based on the multivariate logistic distribution; (2) constructing a set of $m - 1$ nested dichotomies to represent the m categories of the polytomy; and (3) fitting the proportional-odds model to a polytomous response variable with ordered categories.

The three approaches to modeling polytomous data—the polytomous logit model, logit models for nested dichotomies, and the proportional-odds model—address different sets of log-odds, corresponding to different dichotomies constructed from the polytomy. Consider, for example, the ordered polytomy $\{1, 2, 3, 4\}$—representing, say, four ordered educational categories:

- Treating Category 4 as the baseline, the coefficients of the polytomous logit model apply *directly* to the dichotomies $\{1, 4\}$, $\{2, 4\}$, and $\{3, 4\}$ and *indirectly* to any pair of categories.

[52] Consequently, if you are working with software that does not compute the unconstrained-slopes model for cumulative logits, it is generally safe to use the polytomous logit model to formulate an approximate likelihood-ratio test for proportional odds. There is also a score test and a Wald test for the proportional-odds assumption (discussed, e.g., in Long, 1997, sect. 5.5).

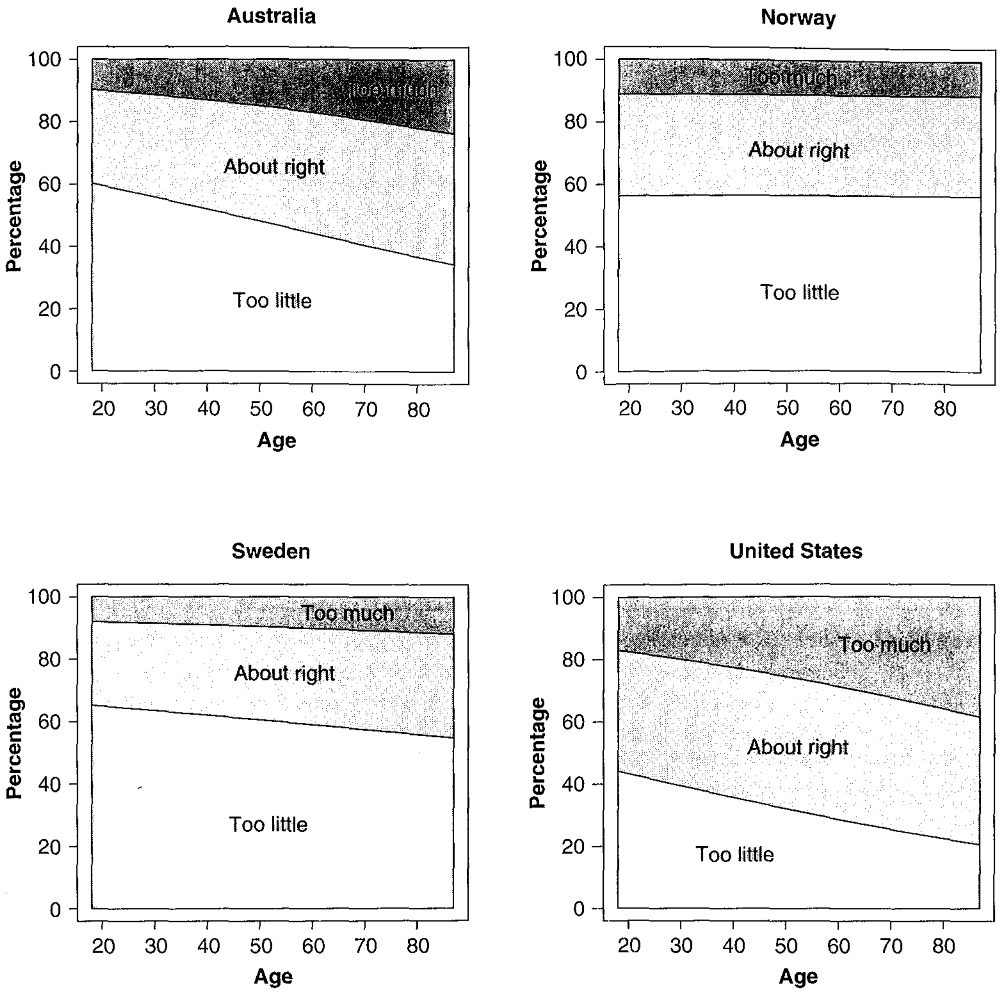

Figure 14.11 Effect display for the interaction of age with country in the proportional-odds model fit to the World Values Survey data. The response variable is assessment of government action for people in poverty.

- Forming continuation dichotomies (one of several possibilities), the nested-dichotomies approach models $\{1, 234\}$, $\{2, 34\}$, and $\{3, 4\}$.
- The proportional-odds model applies to the cumulative dichotomies $\{1, 234\}$, $\{12, 34\}$, and $\{123, 4\}$, imposing the restriction that only the intercepts of the three regression equations differ.

Which of these models is most appropriate depends partly on the structure of the data and partly on our interest in them. If it fits well, the proportional-odds model would generally be preferred for an ordered response on grounds of parsimony, but this model imposes strong structure on the data and may not fit well. Nested dichotomies should only be used if the particular choice of dichotomies makes compelling substantive sense for the data at hand. The implication, then, is that of these three models, the polytomous logit model has the greatest general range of application.[53]

[53] But see Exercise 14.12.

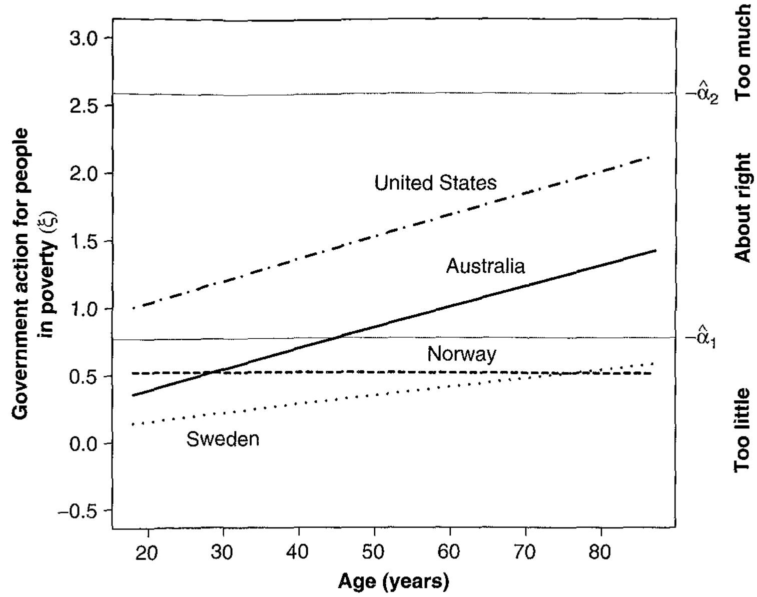

Figure 14.12 Alternative effect display for the proportional-odds model fit to the World Value Survey data, showing fitted values of the latent response. Intercategory thresholds and the corresponding response categories are given at the right of the graph and by the lighter horizontal lines.

14.3 Discrete Explanatory Variables and Contingency Tables

When the explanatory variables—as well as the response—are discrete, the joint sample distribution of the variables defines a contingency table of counts: Each cell of the table records the number of observations possessing a particular combination of characteristics. An example, drawn from *The American Voter* (Campbell, Converse, Miller, & Stokes, 1960), a classical study of electoral behavior, appears in Table 14.9. This table, based on data from sample surveys conducted during the 1956 U.S. presidential election campaign and after the election, relates voting turnout in the election to strength of partisan preference (classified as weak, medium, or strong) and perceived closeness of the election (one-sided or close).

The last column of Table 14.9 gives the *empirical logit* for the response variable,

$$\log_e \frac{\text{Proportion voting}}{\text{Proportion not voting}}$$

for each of the six combinations of categories of the explanatory variables.[54] For example,

$$\text{logit (voted} \mid \text{one-sided, weak preference)} = \log_e \frac{91/130}{39/130} = \log_e \frac{91}{39} = 0.847$$

[54] This calculation will fail if there is a 0 frequency in the table because, in this event, the proportion voting or not voting for some combination of explanatory-variable values will be 0. A simple remedy is to add 0.5 to each of the cell frequencies. Adding 0.5 to each count also serves to reduce the bias of the sample logit as an estimator of the corresponding population logit. See Cox and Snell (1989, pp. 31–32).

Table 14.9 Voter Turnout by Perceived Closeness of the Election and Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

		Turnout		*Logit*
Perceived Closeness	*Intensity of Preference*	*Voted*	*Did Not Vote*	$\log_e \frac{\textit{Voted}}{\textit{Did Not Vote}}$
One-sided	Weak	91	39	0.847
	Medium	121	49	0.904
	Strong	64	24	0.981
Close	Weak	214	87	0.900
	Medium	284	76	1.318
	Strong	201	25	2.084

NOTE: Frequency counts are shown in the body of the table.

Because the conditional proportions voting and not voting share the same denominator, the empirical logit can also be written as

$$\log_e \frac{\text{Number voting}}{\text{Number not voting}}$$

The empirical logits from Table 14.9 are graphed in Figure 14.13, much in the manner of profiles of cell means for a two-way ANOVA.[55] Perceived closeness of the election and intensity of preference appear to interact in affecting turnout: Turnout increases with increasing intensity of preference, but only if the election is perceived to be close. Those with medium or strong preference who perceive the election to be close are more likely to vote than those who perceive the election to be one-sided; this difference is greater among those with strong partisan preference than those with medium partisan preference.

The methods of this chapter are fully appropriate for tabular data. When, as in Table 14.9, the explanatory variables are qualitative or ordinal, it is natural to use logit or probit models that are analogous to ANOVA models. Treating perceived closeness of the election as the "row" factor and intensity of partisan preference as the "column" factor, for example, yields the model

$$\text{logit}\,\pi_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk} \tag{14.30}$$

where

- π_{jk} is the conditional probability of voting in combination of categories j of perceived closeness and k of preference (i.e., in cell jk of the explanatory-variable table);
- μ is the general level of turnout in the population;
- α_j is the main effect on turnout of membership in the jth category of perceived closeness;
- β_k is the main effect on turnout of membership in the kth category of preference; and
- γ_{jk} is the interaction effect on turnout of simultaneous membership in categories j of perceived closeness and k of preference.

[55] See Section 8.2.1.

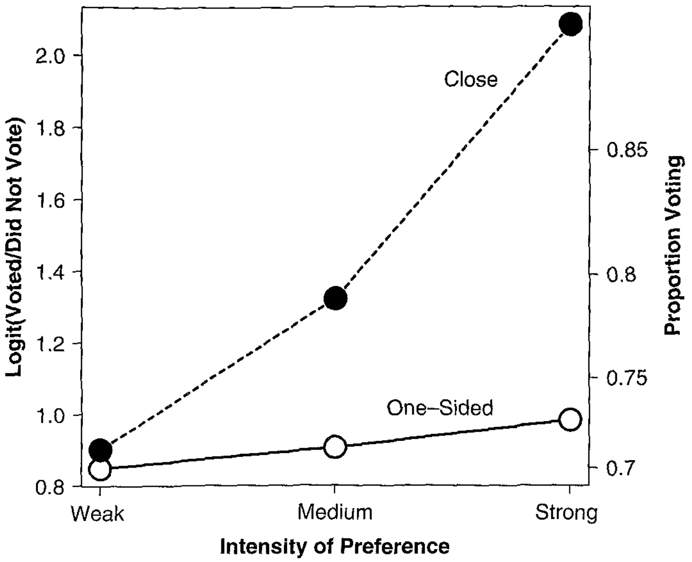

Figure 14.13 Empirical logits for voter turnout by intensity of partisan preference and perceived closeness of the election, for the 1956 U.S. presidential election.

When all the variables—explanatory as well as response—are discrete, their joint distribution defines a contingency table of frequency counts. It is natural to employ logit models that are analogous to ANOVA models to analyze contingency tables.

Under the usual sigma constraints, Equation 14.30 leads to deviation-coded regressors, as in ANOVA. Adapting the SS(·) notation of Chapter 8,[56] likelihood-ratio tests for main effects and interactions can then be constructed in close analogy to the incremental F-tests for the two-way ANOVA model. Residual deviances under several models for the *American Voter* data are shown in Table 14.3, and the analysis-of-deviance table for these data is given in Table 14.11. The log-likelihood-ratio statistic for testing H_0: all $\gamma_{jk} = 0$, for example, is

$$\begin{aligned} G_0^2(\gamma|\alpha,\beta) &= G^2(\alpha,\beta) - G^2(\alpha,\beta,\gamma) \\ &= 1363.552 - 1356.434 \\ &= 7.118 \end{aligned}$$

with $6-4=2$ degrees of freedom, for which $p = .028$. The interaction discerned in Figure 14.13 is, therefore, statistically significant, but not overwhelmingly so.

14.3.1 The Binomial Logit Model*

Although the models for dichotomous and polytomous response variables described in this chapter can be directly applied to tabular data, there is some advantage in reformulating these models to take direct account of the replication of combinations of explanatory-variable values.

[56] In Chapter 8, we used SS(·) to denote the *regression* sum of squares for a model including certain terms. Because the deviance is analogous to the *residual* sum of squares, we need to take differences of deviances in the opposite order.

Table 14.10 Residual Deviances for Models Fit to the *American Voter* Data. Terms: α, Perceived Closeness; β, Intensity of Preference; γ, Closeness × Preference Interaction

Model	*Terms*	*k+1*	*Deviance: G^2*
1	α, β, γ	6	1356.434
2	α, β	4	1363.552
3	α, γ	4	1368.042
4	β, γ	5	1368.554
5	α	2	1382.658
6	β	3	1371.838

NOTE: The column labeled *k*+1 gives the number of parameters in the model, including the constant μ.

Table 14.11 Analysis-of-Deviance Table for the *American Voter* Data

Source	*Models Contrasted*	*df*	G_0^2	*p*
Perceived closeness		1		
$\alpha\|\beta$	6–2		8.286	.0040
$\alpha\|\beta, \gamma$	4–1		12.120	.0005
Intensity of preference		2		
$\beta\|\alpha$	5–2		19.106	<.0001
$\beta\|\alpha, \gamma$	3–1		11.608	.0030
Closeness × Preference		2		
$\gamma\|\alpha, \beta$	2–1		7.118	.028

NOTE: The table shows alternative likelihood-ratio tests for the main effects of perceived closeness of the election and intensity of partisan preference.

In analyzing dichotomous data, for example, we previously treated each observation individually, so the dummy response variable Y_i takes on either the value 0 or the value 1.

Suppose, instead, that we group all the n_i observations that share the specific combination of explanatory-variable values $\mathbf{x}_i' = [x_{i1}, x_{i2}, \dots, x_{ik}]$. Let Y_i count the number of these observations that fall in the first of the two categories of the response variable; we arbitrarily term these observations *successes*. The count Y_i can take on any integer value between 0 and n_i. Let m denote the number of *distinct combinations* of the explanatory variables (e.g., $m = 6$ in Table 14.9 on page 371).

As in our previous development of the dichotomous logit model, let π_i represent $\Pr(\text{success}|\mathbf{x}_i)$. Then the success count Y_i follows the binomial distribution:

$$\begin{aligned} p(y_i) &= \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \qquad (14.31) \\ &= \binom{n_i}{y_i} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i} \end{aligned}$$

To distinguish grouped dichotomous data from ungrouped data, I will refer to the former as *binomial data* and the latter as *binary data*.[57]

[57] Binary data can be thought of as a limiting case of binomial data, for which all $n_i = 1$.

Suppose, next, that the dependence of the response probabilities π_i on the explanatory variables is well described by the logit model

$$\log_e \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i'\boldsymbol{\beta}$$

Substituting this model into Equation 14.31, the likelihood for the parameters is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \binom{n_i}{y_i} [\exp(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} \right)^{n_i}$$

Maximizing the likelihood leads to precisely the same maximum-likelihood estimates, coefficient standard errors, and statistical tests as the binary logit model of Section 14.1.5.[58] The binomial logit model nevertheless has the following advantages:

- Because we deal with m binomial observations rather than the larger $n = \sum_{i=1}^{m} n_i$ binary observations, computations for the binomial logit model are more efficient, especially when the n_i are large.
- The overall residual deviance for the binomial logit model, $-2\log_e L(\mathbf{b})$, implicitly contrasts the model with a *saturated* model that has one parameter for each of the m combinations of explanatory-variable values (e.g., the full two-way "ANOVA" model with main effects and interactions fit in the previous section to the *American Voter* data). The saturated model necessarily recovers the m empirical logits perfectly and, consequently, has a likelihood of 1 and a log likelihood of 0. The residual deviance for a less-than-saturated model, therefore, provides a likelihood-ratio test, on $m-k-1$ degrees of freedom, of the hypothesis that the functional form of the model is correct.[59] In contrast, the residual deviance for the binary logit model cannot be used for a statistical test because the residual degrees of freedom $n-k-1$ (unlike $m-k-1$) grow as the sample size n grows.
- As long as the frequencies n_i are not very small, many diagnostics are much better behaved for the cells of the binomial logit model than for individual binary observations. For example, the individual components of the deviance for the binomial logit model,

$$G_i \equiv \pm\sqrt{-2\left[Y_i \log_e \frac{n_i P_i}{Y_i} + (n_i - Y_i)\log_e \frac{n_i(1-P_i)}{n_i - Y_i}\right]}$$

  can be compared with the unit-normal distribution to locate outlying cells. Here $P_i = 1/[1+\exp(-\mathbf{x}_i'\mathbf{b})]$ is the fitted probability of "success" for cell i, and, therefore, $\widehat{Y}_i = n_i P_i$ is the expected number of "successes" in this cell. The sign of G_i is selected to agree with that of the simple cell residual, $E_i = Y_i - \widehat{Y}_i$.[60]

> Although the binary logit model can be applied to tables in which the response variable is dichotomous, it is also possible to use the equivalent binomial logit model; the binomial logit model is based on the frequency counts of "successes" and "failures" for each combination of explanatory-variable values. When it is applicable, the binomial logit model offers several advantages, including efficient computation, a test of the fit of the model based on its residual deviance, and better-behaved diagnostics.

[58] See Exercise 14.13.

[59] This test is analogous to the test for "lack of fit" in a linear model with a discrete explanatory variable described in Section 12.4.

[60] Diagnostics for logit models and other generalized linear models are discussed in Section 15.4.

Polytomous data can be handled in a similar manner, employing the multinomial distribution.[61] Consequently, all the logit and probit models discussed in this chapter have generalizations to data in which there are repeated observations for combinations of values of the explanatory variables. For example, the *multinomial logit model* generalizes the polytomous logit model of Equation 14.18 (on page 356); indeed, even when it is fit to individual observations, the polytomous logit model is often called the "multinomial logit model" (as I previously mentioned).[62]

Exercises

Exercise 14.1. Nonconstant error variance in the linear-probability model: Make a table showing the variance of the error $V(\varepsilon) = \pi(1-\pi)$ for the following values of π:

$$.001,\ .01,\ .05,\ .1,\ .3,\ .5.,\ .7,\ .9,\ .95,\ .99,\ .999$$

When is the heteroscedasticity problem serious?

Exercise 14.2. Show that using the cumulative rectangular distribution as $P(\cdot)$ in the general model

$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i)$$

produces the constrained linear-probability model. (See Section 14.1.2.)

Exercise 14.3. *Show that the slope of the logistic-regression curve, $\pi = 1/\left[1 + e^{-(\alpha+\beta X)}\right]$, can be written as $\beta\pi(1-\pi)$. (*Hint*: Differentiate π with respect to X, and then substitute for expressions that equal π and $1-\pi$.)

Exercise 14.4. Substitute first $y_i = 0$ and then $y_i = 1$ into the expression

$$p(y_i) \equiv \Pr(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

to show that this equation captures $p(0) = 1 - \pi_i$ and $p(1) = \pi_i$.

Exercise 14.5. *Show that, for the logit multiple-regression model,

$$\pi_i = \frac{1}{1 + \exp[-(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})]}$$

the probability that $Y_i = 0$ can be written as

$$1 - \pi_i = \frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}$$

Exercise 14.6. *Show that the maximized likelihood for the fitted logit model can be written as

$$\log_e L = \sum_{i=1}^{n}\left[y_i \log_e P_i + (1-y_i)\log_e(1-P_i)\right]$$

[61] See Exercise 14.14.

[62] As in the case of binary data, we can think of individual polytomous observations as multinomial observations in which all the total counts are $n_i = 1$.

where

$$P_i = \frac{1}{1 + \exp[-(A + B_1 X_{i1} + \cdots + B_k X_{ik})]}$$

is the fitted probability that $Y_i = 1$. [*Hint*: Use $p(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$.]

Exercise 14.7. *Residual deviance in least-squares regression: The log likelihood for the linear regression model with normal errors can be written as

$$\log_e L(\alpha, \beta_1, \ldots, \beta_k, \sigma_\varepsilon^2) = -\frac{n}{2}\log_e\left(2\pi\sigma_\varepsilon^2\right) - \frac{\sum_{i=1}^n \varepsilon_i^2}{2\sigma_\varepsilon^2}$$

where $\varepsilon_i = Y_i - (\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$ (see Section 9.3.3). Let l represent the maximized log likelihood, treated as a function of the regression coefficients $\alpha, \beta_1, \ldots, \beta_k$ but not of the error variance σ_ε^2, which is regarded as a "nuisance parameter." Let $l' = -(n/2)\log_e(2\pi\sigma_\varepsilon^2)$ represent the log likelihood for a model that fits the data perfectly (i.e., for which all $\varepsilon_i = 0$). Then the residual deviance is defined as $-2\sigma_\varepsilon^2(l - l')$. Show that, by this definition, the residual deviance for the normal linear model is just the residual sum of squares. (For the logit model, there is no nuisance parameter, and $l' = 0$; the residual deviance for this model is, therefore, $-2\log_e L$, as stated in the text. See Chapter 15 for further discussion of the deviance.)

Exercise 14.8. *Evaluate the information matrix for the logit model,

$$\mathcal{J}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]$$

and show that the estimated asymptotic covariance matrix of the coefficients is

$$\widehat{\mathcal{V}}(\mathbf{b}) = \left[\sum_{i=1}^n \frac{\exp(-\mathbf{x}_i'\mathbf{b})}{[1 + \exp(-\mathbf{x}_i'\mathbf{b})]^2}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$$

Exercise 14.9. *Show that the maximum-likelihood estimator for the logit model can be written as

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}^*$$

where

$$\mathbf{y}^* \equiv \mathbf{X}\mathbf{b} + \mathbf{V}^{-1}(\mathbf{y} - \mathbf{p})$$

(*Hint*: Simply multiply out the equation.)

Exercise 14.10. *Show that the polytomous logit model of Equation 14.18 (page 356) can be written in the form

$$\log_e \frac{\pi_{ij}}{\pi_{im}} = \gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik} \quad \text{for } j = 1, \ldots, m-1$$

Exercise 14.11. *Derive the estimating equations (Equation 14.21 on page 360) and the information matrix (Equations 14.22 and 14.23) for the polytomous logit model.

Exercise 14.12. Independence From Irrelevant Alternatives: In the polytomous logit model discussed in Section 14.2.1, the logit for a particular pair of categories depends on the coefficients for those categories but not on those for other categories in the model. Show that this is the case. (*Hint*: See Equation 14.2.1.) In the context of a discrete-choice model (e.g., Greene, 2003, chap. 21; or Alvarez & Nagler, 1998), this property can be interpreted to mean that the relative odds for a pair of categories is independent of the other categories in the choice set. Why is this often an implausible assumption? (*Hint:* Consider a multiparty election in a jurisdiction, such as Canada or the U.K., where some parties field candidates in only part of the country, or what happens to the electoral map when a new party is formed.) For this reason, models such as the polytomous probit model that *do not* assume independence from irrelevant alternatives are sometimes preferred.

Exercise 14.13. *Derive the maximum-likelihood estimating equations for the binomial logit model. Show that this model produces the same estimated coefficients as the dichotomous (binary) logit model of Section 15.1. (*Hint*: Compare the log likelihood for the binomial model with the log likelihood for the binary model; by separating individual observations sharing a common set of X values, show that the former log likelihood is equal to the latter, except for a constant factor. This constant is irrelevant because it does not influence the maximum-likelihood estimator; moreover, the constant disappears in likelihood-ratio tests.)

Exercise 14.14. *Use the multinomial distribution (see Appendix D) to specify a polytomous logit model for discrete explanatory variables (analogous to the binomial logit model), where combinations of explanatory-variable values are replicated. Derive the likelihood under the model, and the maximum-likelihood estimating equations.

Summary

- It is problematic to apply least-squares linear regression to a dichotomous response variable: The errors cannot be normally distributed and cannot have constant variance. Even more fundamentally, the linear specification does not confine the probability for the response to the unit interval.
- More adequate specifications transform the linear predictor $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ smoothly to the unit interval, using a cumulative probability distribution function $P(\cdot)$. Two such specifications are the probit and the logit models, which use the normal and logistic CDFs, respectively. Although these models are very similar, the logit model is simpler to interpret because it can be written as a linear model for the log-odds:

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

- The dichotomous logit model can be fit to data by the method of maximum likelihood. Wald tests and likelihood-ratio tests for the coefficients of the model parallel t-tests and incremental F-tests for the general linear model. The residual deviance for the model, defined as $G^2 = -2 \times$ the maximized log likelihood, is analogous to the residual sum of squares for a linear model.
- Several approaches can be taken to modeling polytomous data, including:

  1. modeling the polytomy directly using a logit model based on the multivariate logistic distribution;
  2. constructing a set of $m - 1$ nested dichotomies to represent the m categories of the polytomy; and

3. fitting the proportional-odds model to a polytomous response variable with ordered categories.

- When all the variables—explanatory as well as response—are discrete, their joint distribution defines a contingency table of frequency counts. It is natural to employ logit models that are analogous to ANOVA models to analyze contingency tables. Although the binary logit model can be applied to tables in which the response variable is dichotomous, it is also possible to use the equivalent binomial logit model; the binomial logit model is based on the frequency counts of "successes" and "failures" for each combination of explanatory-variable values. When it is applicable, the binomial logit model offers several advantages, including efficient computation, a test of the fit of the model based on its residual deviance, and better-behaved diagnostics. There are analogous logit and probit models, such as the multinomial logit model, for polytomous responses.

Recommended Reading

The topics introduced in this chapter could easily be expanded to fill several books, and there is a large literature—both in journals and texts—dealing with logit and related models for categorical response variables, and with the analysis of contingency tables.[63]

- Agresti (2002) presents an excellent and comprehensive overview of statistical methods for qualitative data. The emphasis is on logit and log-linear models for contingency tables, but there is some consideration of logistic regression models and other topics. Also see Agresti (1996) for a briefer and lower-level treatment of much of this material.
- Fienberg's (1980) widely read text on the analysis of contingency tables provides an accessible and lucid introduction to log-linear models and related subjects, such as logit models and models for ordered categories.
- The second edition of Cox and Snell's (1989) classic text concentrates on logit models for dichotomous data but also includes some discussion of polytomous nominal and ordinal data.
- Collett (2003) also focuses on the binary and binomial logit models. The book is noteworthy for its extensive review of diagnostic methods for logit models.
- Greene (2003, chap. 21) includes a broad treatment of models for categorical responses from the point of view of "discrete choice models" in econometrics.
- Long (1997) and Powers and Xie (2000) both present high-quality, accessible expositions for social scientists of statistical models for categorical data.

[63] Also see the references on generalized linear models given at the end of the next chapter, which briefly describes log-linear models for contingency tables.

15 Generalized Linear Models

Due originally to Nelder and Wedderburn (1972), generalized linear models are a remarkable synthesis and extension of familiar regression models such as the linear models described in Part II of this text and the logit and probit models described in the preceding chapter. The current chapter begins with a consideration of the general structure and range of application of generalized linear models; proceeds to examine in greater detail generalized linear models for count data, including contingency tables; briefly sketches the statistical theory underlying generalized linear models; and concludes with the extension of regression diagnostics to generalized linear models.

The unstarred sections of this chapter are perhaps more difficult than the unstarred material in preceding chapters. Generalized linear models have become so central to effective statistical data analysis, however, that it is worth the additional effort required to acquire a basic understanding of the subject.

15.1 The Structure of Generalized Linear Models

A *generalized linear model* (or GLM[1]) consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In Nelder and Wedderburn's original formulation, the distribution of Y_i is a member of an *exponential family*, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions. Subsequent work, however, has extended GLMs to multivariate exponential families (such as the multinomial distribution), to certain non-exponential families (such as the two-parameter negative-binomial distribution), and to some situations in which the distribution of Y_i is not specified completely. Most of these ideas are developed later in the chapter.
2. A *linear predictor*—that is a linear function of regressors

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

As in the linear model, and in the logit and probit models of Chapter 14, the regressors X_{ij} are prespecified functions of the explanatory variables and therefore may include quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial regressors, dummy regressors, interactions, and so on. Indeed, one of the advantages of GLMs is that the structure of the linear predictor is the familiar structure of a linear model.

3. A smooth and invertible linearizing *link function* $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

[1] Some authors use the acronym "GLM" to refer to the "*general* linear model"—that is, the linear regression model with normal errors described in Part II of the text—and instead employ "GLIM" to denote *generalized* linear models (which is also the name of a computer program used to fit GLMs).