transformation $\log[\pi/(1-\pi)]$ of π, symbolized by $\text{logit}(\pi)$. Logistic regression models are often called *logit models*. The logit is the natural parameter of the binomial distribution, so the logit link is its canonical link. Whereas π is restricted to the range $(0, 1)$, the logit can be any real number. The real numbers are also the potential range for linear predictors (such as $\alpha + \beta x$) that form the systematic component of a GLM, so this model does not have the structural problem that the linear probability model has.

The parameter β in (4.2.2) determines the rate of increase or decrease of the curve. When $\beta > 0$, $\pi(x)$ increases as x increases, as in Figure 4.2*a*. When $\beta < 0$, $\pi(x)$ decreases as x increases, as in Figure 4.2*b*. The magnitude of β determines how fast the curve increases or decreases. As $|\beta|$ increases, the curve has a steeper rate of change. When $\beta = 0$, the curve flattens to a horizontal straight line.

For the snoring and heart disease data in Table 4.1, software reports the ML fit for the logistic regression model of

$$\text{logit}[\hat{\pi}(x)] = -3.87 + 0.40x.$$

The positive value of $\hat{\beta} = 0.40$ reflects the increased chance of heart disease at higher levels of snoring. Chapter 5 presents several ways of interpreting such equations. For instance, it shows how to calculate the predicted probabilities for the model fit. Table 4.1 also reports these fitted values, and Figure 4.1 displays the fit. The fit is close to linear over this rather narrow range of predicted probabilities, and results are similar to those for the linear probability model.

4.2.4 Alternative Binary Links*

For the logistic regression curves pictured in Figure 4.2, the probability of a success increases continuously or decreases continuously as x increases. Let X denote a random variable, and let x denote a potential value for X. The cumulative distribution function (*cdf*) $F(x)$ for X is defined as

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

Such a function, plotted as a function of x, has appearance like that in Figure 4.2a. As x increases, $F(x)$ increases gradually from 0 to 1, since $P(X \leq x)$ increases as x increases. This suggests a class of models for binary responses whereby the dependence of $\pi(x)$ on x has form

$$\pi(x) = F(x), \tag{4.2.3}$$

where F is a *cdf* for some distribution.

The logistic regression curve has this form. When $\beta > 0$, $F(x)$ is the *cdf* of a two-parameter *logistic distribution*. When $\beta < 0$, the formula for $1 - \pi(x)$ has the logistic *cdf* appearance. Each choice of α and of $\beta > 0$ corresponds to a different logistic distribution. The logistic *cdf* corresponds to a probability distribution with

a symmetric, bell shape. In fact, it looks similar to a normal distribution but with slightly thicker tails.

Model form (4.2.3) occurs naturally when a *tolerance distribution* applies to subjects' responses. For instance, in a toxicology study, suppose that researchers spray an insectide at various dosage levels on batches of mosquitoes. For each mosquito, the response is whether it dies. Each mosquito may have a certain tolerance to the insecticide, such that it dies if the dosage level exceeds its tolerance and survives if the dosage level is less than its tolerance. Tolerances would vary among mosquitoes. If a *cdf* F describes the distribution of tolerances, then the model for the probability $\pi(x)$ of death at dosage level x necessarily has form (4.2.3). For instance, if the tolerances vary among mosquitoes according to a logistic distribution, then the logistic regression model applies.

4.2.5 Probit Models*

When F is the *cdf* of a normal distribution, model type (4.2.3) is called the *probit model*. The link function for the model is then called the *probit link*. The probit model has alternative expression

$$\text{probit}[\pi(x)] = \alpha + \beta x.$$

The probit link applied to a probability $\pi(x)$ transforms it to the standard normal z-score at which the left-tail probability equals $\pi(x)$. For instance, probit(.05) $= -1.645$, probit(.50) $= 0$, probit(.95) $= 1.645$, and probit(.975) $= 1.96$. The probit model is a GLM with binomial random component and probit link.

We illustrate using the snoring and heart disease data. The ML fit of the probit model, using scores $\{0, 2, 4, 5\}$ for snoring level, is

$$\text{probit}[\hat{\pi}(x)] = -2.061 + 0.188x.$$

At snoring level $x = 0$, the fitted probit equals $-2.061 + 0.188(0) = -2.06$. The fitted probability $\hat{\pi}(0)$ is the left-tail probability for the standard normal distribution at -2.06, which equals .020. At snoring level $x = 5$, the fitted probit equals $-2.061 + 0.188(5) = -1.12$, which corresponds to a fitted probability of .131.

The fitted values, shown in Table 4.1 and Figure 4.1, are similar to those obtained with the linear probability and logistic regression models. For practical purposes, probit and logistic regression curves look the same. It is rare, and requires enormous sample sizes, to find data for which a logistic regression model fits well but the probit model fits poorly, or conversely. Parameter estimates differ for the two models, since their links have different scales. When both models fit well, slope estimates in logistic regression models are roughly about 1.6–2.0 times those in probit models.

The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1 - \pi(x)$, when $\beta < 0$) has the appearance of the normal *cdf* with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$. For the snoring and heart disease data, the probit fit corresponds to a normal *cdf* having mean of $-\hat{\alpha}/\hat{\beta} = 2.061/0.188 = 11.0$ and standard deviation of $1/|\hat{\beta}| = 1/0.188 = 5.3$. The predicted probability of heart

disease equals $\frac{1}{2}$ at snoring level $x = 11.0$; that is, $x = 11.0$ has a fitted probit of $-2.061 + 0.188(11) = 0$, which is the z-score corresponding to a left-tail probability of $\frac{1}{2}$. The fitted probit value of -2.06 at $x = 0$ means that 0 is 2.06 standard deviations below the mean of a normal distribution with mean 11.0 and standard deviation 5.3. Since snoring level is restricted to the range 0–5 for these data, well below 11, the fitted probabilities over this range are quite small.

The probit model was introduced in 1934 for models in toxicology. The logistic regression model was not studied until about a decade later, but it is now much more popular than the probit. Partly this is because one can also interpret the logistic regression effects using odds ratios. Thus, one can fit those models to data from case-control studies, because one can estimate odds ratios for such data (Sections 2.3.4, 5.1.4, 9.2.3).

4.3 GENERALIZED LINEAR MODELS FOR COUNT DATA: POISSON REGRESSION

Many discrete response variables have counts as possible outcomes. For instance, for a sample of cities worldwide, each observation might be the number of automobile thefts in 1995. Or, for a sample of silicon wafers used in manufacturing computer chips, each observation might be the number of imperfections on a wafer. This section introduces GLMs for count data. These GLMs assume a Poisson distribution for the random component. Like counts, Poisson variates can take any nonnegative integer value.

Section 1.2 introduced the Poisson distribution as a sampling model for counts. Chapter 6 presents Poisson GLMs for counts in contingency tables. The response data are cell counts obtained by cross-classifying subjects on two or more categorical response variables. This section introduces Poisson regression-type models using an alternative application: modeling count or rate data for a single response variable.

4.3.1 Poisson Regression

The Poisson distribution has a positive mean. Though one can model the Poisson mean in GLMs using the identity link, it is more common to model the log of the mean. Like the linear predictor $\alpha + \beta x$, the log of the mean can take any real value. The log mean is the natural parameter for the Poisson distribution, and the log link is the canonical link for a GLM with Poisson random component. A *Poisson loglinear model* is a GLM that assumes a Poisson distribution for Y and uses the log link.

Let μ denote the expected value for a Poisson variate Y, and let X denote an explanatory variable. The Poisson loglinear model has form

$$\log \mu = \alpha + \beta x. \tag{4.3.1}$$

For this model, the mean satisfies the exponential relationship

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}. \tag{4.3.2}$$