```
female         -0.01      0.15
black          -3.64      0.59
income         -0.03      0.07

glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1972))
            coef.est coef.se
(Intercept)  0.67     0.18
female      -0.25     0.12
black       -2.63     0.27
income       0.09     0.05
```

What happened with the coefficient of `black` in 1964? Take a look at the data and figure out where this extreme estimate came from. What can be done to fit the model in 1964?

CHAPTER 6

Generalized linear models

6.1 Introduction

Generalized linear modeling is a framework for statistical analysis that includes linear and logistic regression as special cases. Linear regression directly predicts continuous data y from a *linear predictor* $X\beta = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$. Logistic regression predicts $\Pr(y=1)$ for binary data from a linear predictor with an inverse-logit transformation. A generalized linear model involves:

1. A data vector $y = (y_1, \ldots, y_n)$
2. Predictors X and coefficients β, forming a linear predictor $X\beta$
3. A *link function* g, yielding a vector of transformed data $\hat{y} = g^{-1}(X\beta)$ that are used to model the data
4. A data distribution, $p(y|\hat{y})$
5. Possibly other parameters, such as variances, overdispersions, and cutpoints, involved in the predictors, link function, and data distribution.

The options in a generalized linear model are the transformation g and the data distribution p.

- In *linear regression*, the transformation is the identity (that is, $g(u) \equiv u$) and the data distribution is normal, with standard deviation σ estimated from data.
- In *logistic regression*, the transformation is the inverse-logit, $g^{-1}(u) = \text{logit}^{-1}(u)$ (see Figure 5.2a on page 80) and the data distribution is defined by the probability for binary data: $\Pr(y=1) = \hat{y}$.

This chapter discusses several other classes of generalized linear model, which we list here for convenience:

- The *Poisson* model (Section 6.2) is used for count data; that is, where each data point y_i can equal 0, 1, 2, ... The usual transformation g used here is the logarithmic, so that $g(u) = \exp(u)$ transforms a continuous linear predictor $X_i\beta$ to a positive $\hat{y}_i$. The data distribution is Poisson.

  It is usually a good idea to add a parameter to this model to capture *overdispersion*, that is, variation in the data beyond what would be predicted from the Poisson distribution alone.
- The *logistic-binomial* model (Section 6.3) is used in settings where each data point y_i represents the number of successes in some number n_i of tries. (This n_i, the number of tries for data point i, is not the same as n, the number of data points.) In this model, the transformation g is the inverse-logit and the data distribution is binomial.

  As with Poisson regression, the binomial model is typically improved by the inclusion of an overdispersion parameter.
- The *probit* model (Section 6.4) is the same as logistic regression but with the logit function replaced by the normal cumulative distribution, or equivalently with the normal distribution instead of the logistic in the latent-data errors.

- *Multinomial* logit and probit models (Section 6.5) are extensions of logistic and probit regressions for categorical data with more than two options, for example survey responses such as Strongly Agree, Agree, Indifferent, Disagree, Strongly Disagree. These models use the logit or probit transformation and the multinomial distribution and require additional parameters to model the multiple possibilities of the data.

  Multinomial models are further classified as *ordered* (for example, Strongly Agree, . . . , Strongly Disagree) or *unordered* (for example, Vanilla, Chocolate, Strawberry, Other).
- *Robust* regression models (Section 6.6) replace the usual normal or logistic models by other distributions[1] (usually the so-called Student-t family of models) that allow occasional extreme values.

This chapter briefly goes through many of these models, with an example of overdispersed Poisson regression in Section 6.2 and an ordered logistic example in Section 6.5. Finally, in Section 6.8 we discuss the connections between generalized linear models and behavioral models of choice that are used in psychology and economics, using as an example the logistic regression for well switching in Bangladesh. The chapter is not intended to be a comprehensive overview of generalized linear models; rather, we want to give a sense of the variety of regression models that can be appropriate for different data structures that we have seen in applications.

Fitting generalized linear models in R

Because of the variety of options involved, generalized linear modeling can be more complicated than linear and logistic regressions. The starting point in R is the `glm()` function, which we have already used extensively for logistic regression in Chapter 5 and is a generalization of the linear-modeling function `lm()`. We can use `glm()` directly to fit logistic-binomial, probit, and Poisson regressions, among others, and to correct for overdispersion where appropriate, with `bayesglm()` available when there is separation or the model is otherwise nonidentified. Ordered logit and probit regressions can be fit using the `polr()` and `bayespolr()` functions, unordered probit models can be fit using the `mnp` package, and t models can be fit using the `hett` package in R. (See Appendix C for information on these and other R packages.) Beyond this, most of these models and various generalizations can be fit in Bugs, as we discuss in Part 2B of this book in the context of multilevel modeling.

6.2 Poisson regression, exposure, and overdispersion

The Poisson distribution is used to model variation in count data (that is, data that can equal $0, 1, 2, \ldots$). After a brief introduction, we illustrate in detail with the example of New York City police stops that we introduced in Section 1.2.

Traffic accidents

In the Poisson model, each unit i corresponds to a setting (typically a spatial location or a time interval) in which y_i events are observed. For example, i could index street intersections in a city and y_i could be the number of traffic accidents at intersection i in a given year.

[1] In the statistical literature, generalized linear models have been defined using exponential-family models, a particular class of data distributions that excludes, for example, the t distribution. For our purposes, however, we use the term "generalized linear model" to apply to any model with a linear predictor, link function, and data distribution, not restricting to exponential-family models.

As with linear and logistic regression, the variation in y can be explained with linear predictors X. In the traffic accidents example, these predictors could include: a constant term, a measure of the average speed of traffic near the intersection, and an indicator for whether the intersection has a traffic signal. The basic Poisson regression model has the form

$$y_i \sim \text{Poisson}(\theta_i). \tag{6.1}$$

The parameter θ_i must be positive, so it makes sense to fit a linear regression on the logarithmic scale:

$$\theta_i = \exp(X_i\beta). \tag{6.2}$$

Interpreting Poisson regression coefficients

The coefficients β can be exponentiated and treated as multiplicative effects. For example, suppose the traffic accident model is

$$y_i \sim \text{Poisson}(\exp(2.8 + 0.012X_{i1} - 0.20X_{i2})),$$

where X_{i1} is average speed (in miles per hour, or mph) on the nearby streets and $X_{i2} = 1$ if the intersection has a traffic signal or 0 otherwise. We can then interpret each coefficient as follows:

- The constant term gives the intercept of the regression, that is, the prediction if $X_{i1} = 0$ and $X_{i2} = 0$. Since this is not possible (no street will have an average speed of 0), we will not try to interpret the constant term.
- The coefficient of X_{i1} is the expected difference in y (on the logarithmic scale) for each additional mph of traffic speed. Thus, the expected multiplicative increase is $e^{0.012} = 1.012$, or a 1.2% positive difference in the rate of traffic accidents per mph. Since traffic speeds vary by tens of mph, it would actually make sense to define X_{i1} as speed in tens of mph, in which case its coefficient would be 0.12, corresponding to a 12% increase (more precisely, $e^{0.12} = 1.127$: a 12.7% increase) in accident rate per ten mph.
- The coefficient of X_{i2} tells us that the predictive difference of having a traffic signal can be found be multiplying the accident rate by $\exp(-0.20) = 0.82$ yielding a reduction of 18%.

As with regression models in general, each coefficient is interpreted as a comparison in which one predictor differs by one unit while all the other predictors remain at the same level, which is not necessarily the most appropriate assumption when extending the model to new settings. For example, installing traffic signals in all the intersections in the city would *not* necessarily be expected to reduce accidents by 18%.

Poisson regression with an exposure input

In most applications of Poisson regression, the counts can be interpreted relative to some baseline or "exposure," for example, the number of vehicles that travel through the intersection. In the general Poisson regression model, we think of y_i as the number of cases in a process with rate θ_i and exposure u_i.

$$y_i \sim \text{Poisson}(u_i\theta_i), \tag{6.3}$$

where, as before, $\theta_i = \exp(X_i\beta)$. The logarithm of the exposure, $\log(u_i)$, is called the *offset* in generalized linear model terminology.

The regression coefficients β summarize the associations between the predictors and θ_i (in our example, the rate of traffic accidents per vehicle).

Including log(exposure) as a predictor in the Poisson regression. Putting the logarithm of the exposure into the model as an offset, as in model (6.3), is equivalent to including it as a regression predictor, but with its coefficient fixed to the value 1. Another option is to include it as a predictor and let its coefficient be estimated from the data. In some settings, this makes sense in that it can allow the data to be fit better; in other settings, it is simpler to just keep it as an offset so that the estimated rate θ has a more direct interpretation.

Differences between the binomial and Poisson models

The Poisson model is similar to the binomial model for count data (see Section 6.3) but is applied in slightly different situations:

- If each data point y_i can be interpreted as the number of "successes" out of n_i trials, then it is standard to use the binomial/logistic model (as described in Section 6.3) or its overdispersed generalization.
- If each data point y_i does not have a natural limit—it is not based on a number of independent trials—then it is standard to use the Poisson/logarithmic regression model (as described here) or its overdispersed generalization.

Example: police stops by ethnic group

For the analysis of police stops:

- The units i are precincts and ethnic groups ($i = 1, \dots, n = 3 \times 75$).
- The outcome y_i is the number of stops of members of that ethnic group in that precinct.
- The exposure u_i is the number of arrests by people of that ethnic group in that precinct in the previous year as recorded by the Department of Criminal Justice Services (DCJS).
- The inputs are the precinct and ethnicity indexes.
- The predictors are a constant, 74 precinct indicators (for example, precincts 2–75, with precinct 1 as the baseline), and 2 ethnicity indicators (for example, for hispanics and whites, with blacks as the baseline).

We illustrate model fitting in three steps. First, we fit a model with the offset and a constant term alone:

R output

```
glm(formula = stops ~ 1, family=poisson, offset=log(arrests))
                coef.est coef.se
(Intercept)        -3.4     0.0
  n = 225, k = 1
  residual deviance = 44877, null deviance = 44877 (difference = 0)
```

Next, we add ethnicity indicators:

R output

```
glm(formula = stops ~ factor(eth), family=poisson,
    offset=log(arrests))
                coef.est coef.se
(Intercept)        -3.30    0.00
```

```
factor(eth)2       0.06    0.01
factor(eth)3      -0.18    0.01
  n = 225, k = 3
  residual deviance = 44133, null deviance = 44877 (difference = 744.1)
```

The two ethnicity coefficients are highly statistically significant, and the deviance has decreased by 744, much more than the 2 that would be expected if ethnicity had no explanatory power in the model. Compared to the baseline category 1 (blacks), we see that category 2 (hispanics) has 6% more stops, and category 3 (whites) has 18% fewer stops, in proportion to DCJS arrest rates.

Now we add the 75 precincts:

R output

```
glm(formula = stops ~ factor(eth) + factor(precinct), family=poisson,
    offset=log(arrests))
                    coef.est coef.se
(Intercept)           -4.03    0.05
factor(eth)2           0.00    0.01
factor(eth)3          -0.42    0.01
factor(precinct)2     -0.06    0.07
factor(precinct)3      0.54    0.06
. . .
factor(precinct)75     1.41    0.08
  n = 225, k = 77
  residual deviance = 2828.6, null deviance = 44877 (difference = 42048.4)
  overdispersion parameter = 18.2
```

The decrease in the deviance, from 44,000 to 2800, is huge—much larger than the decrease of 74 that would be expected if the precinct factor were random noise. After controlling for precincts, the ethnicity coefficients have changed a bit—blacks and hispanics (categories 1 and 2) have approximately the same rate of being stopped, and whites (category 3) have about a 42% lower chance than minorities of being stopped—all in comparison to the DCJS arrest rates, which are used as a baseline.[2]

Thus, controlling for precinct actually increases the difference between whites and minorities in the rate of stops. We explore this issue further in Section 15.1.

We can also look at the precinct coefficients in the regression—for example, the stop rates (per DCJS arrest) after controlling for ethnicity, are approximately 6% lower in precinct 2, $\exp(0.54) = 1.72$ times as high in precinct 3, ..., and $\exp(1.41) = 4.09$ times as high in precinct 75, as compared to the baseline precinct 1.

The exposure input

In this example, stops by police are compared to the number of arrests in the previous year, so that the coefficient for the "hispanic" or "white" indicator will be greater than 1 if the people in that group are stopped disproportionately to their rates of arrest, as compared to blacks. Similarly, the coefficients for the indicators for precincts 2–75 will exceed 1 for those precincts where stops are more frequent than in precinct 1, as compared to their arrest rates in the previous year.

In Section 15.1 we shall consider another possible analysis that uses population, rather than previous year's arrests, as the exposure.

[2] More precisely, the exponentiated coefficient for whites is $\exp(-0.42) = 0.66$, so their chance of being stopped is actually 34% lower—the approximation $\exp(-\beta) \approx 1-\beta$ is accurate only when β is close to 0.

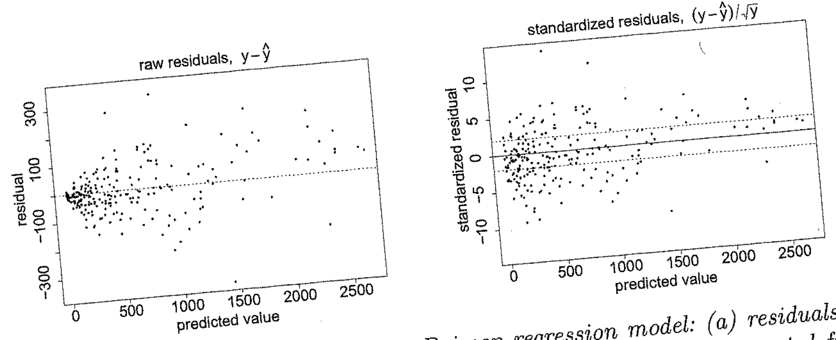

Figure 6.1 *Testing for overdispersion in a Poisson regression model: (a) residuals versus predicted values, (b) standardized residuals versus predicted values. As expected from the model, the variance of the residuals increases as predicted values increase. The standardized residuals should have mean 0 and standard deviation 1 (hence the lines at ±2 indicating approximate 95% error bounds). The variance of the standardized residuals is much greater than 1, indicating a large amount of overdispersion.*

Overdispersion

Poisson regressions do not supply an independent variance parameter σ, and as a result can be overdispersed and usually are, a point we considered briefly on page 21 and pursue further here in a regression context. Under the Poisson distribution the variance equals the mean—that is, the standard deviation equals the square root of the mean. In the model (6.3), $\mathrm{E}(y_i) = u_i\theta_i$ and $\mathrm{sd}(y_i) = \sqrt{u_i\theta_i}$. We define the standardized residuals:

$$\begin{aligned} z_i &= \frac{y_i - \hat{y}_i}{\mathrm{sd}(\hat{y}_i)} \\ &= \frac{y_i - u_i\hat{\theta}_i}{\sqrt{u_i\hat{\theta}_i}}, \end{aligned} \tag{6.4}$$

where $\hat{\theta}_i = e^{X_i\hat{\beta}}$. If the Poisson model is true, then the z_i's should be approximately independent (not exactly independent, since the same estimate $\hat{\beta}$ is used in computing all of them), each with mean 0 and standard deviation 1. If there is overdispersion, however, we would expect the z_i's to be larger, in absolute value, reflecting the extra variation beyond what is predicted under the Poisson model.

We can test for overdispersion in classical Poisson regression by computing the sum of squares of the n standardized residuals, $\sum_{i=1}^n z_i^2$, and comparing this to the χ^2_{n-k} distribution, which is what we would expect under the model (using $n-k$ rather than n degrees of freedom to account for the estimation of k regression coefficients). The χ^2_{n-k} distribution has average value $n-k$, and so the ratio,

$$\text{estimated overdispersion} = \frac{1}{n-k}\sum_{i=1}^{n} z_i^2, \tag{6.5}$$

is a summary of the overdispersion in the data compared to the fitted model.

For example, the classical Poisson regression for the police stops has $n = 225$ data points and $k = 77$ linear predictors. Figure 6.1 plots the residuals $y_i - \hat{y}_i$ and standardized residuals $z_i = (y_i - \hat{y}_i)/\mathrm{sd}(\hat{y}_i)$, as a function of predicted values from the Poisson regression model. As expected from the Poisson model, the variance of the residuals increases as the predicted values increase, and the variance of the standardized residuals is approximately constant. However, the standardized residuals have a variance much greater than 1, indicating serious overdispersion.

To program the overdispersion test in R:

R code

```
yhat <- predict (glm.police, type="response")
z <- (stops-yhat)/sqrt(yhat)
cat ("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
cat ("p-value of overdispersion test is ", pchisq (sum(z^2), n-k), "\n")
```

The sum of squared standardized residuals is $\sum_{i=1}^n z_i^2 = 2700$, compared to an expected value of $n-k = 148$. The estimated overdispersion factor is $2700/148 = 18.2$, and the p-value is 1, indicating that the probability is essentially zero that a random variable from a χ^2_{148} distribution would be as large as 2700. In summary, the police stops data are overdispersed by a factor of 18, which is huge—even an overdispersion factor of 2 would be considered large—and also statistically significant.

Adjusting inferences for overdispersion

In this example, the basic correction for overdispersion is to multiply all regression standard errors by $\sqrt{18.2} = 4.3$. Luckily, it turns out that our main inferences are not seriously affected. The parameter of primary interest is α_3—the log of the rate of stops for whites compared to blacks—which is estimated at -0.42 ± 0.01 before (see the regression display on page 113) and now becomes -0.42 ± 0.04. Transforming back to the original scale, whites are stopped at an estimated 66% of the rate of blacks, with an approximate 50% interval of $e^{-0.42\pm(2/3)0.04} = [0.64, 0.67]$ and an approximate 95% interval of $e^{-0.42\pm 2\cdot 0.04} = [0.61, 0.71]$.

Fitting the overdispersed-Poisson or negative-binomial model

More simply, we can fit an overdispersed model using the quasipoisson family:

R output

```
glm(formula = stops ~ factor(eth) + factor(precinct), family=quasipoisson,
    offset=log(arrests))
                   coef.est coef.se
(Intercept)          -4.03    0.21
factor(eth)2          0.00    0.03
factor(eth)3         -0.42    0.04
factor(precinct)2    -0.06    0.30
factor(precinct)3     0.54    0.24
. . .
factor(precinct)75    1.41    0.33
  n = 225, k = 77
  residual deviance = 2828.6, null deviance = 44877 (difference = 42048.4)
  overdispersion parameter = 18.2
```

We write this model as

$$y_i \sim \text{overdispersed Poisson}\,(u_i \exp(X_i\beta),\, \omega),$$

where ω is the overdispersion parameter (estimated at 18.2 in this case). Strictly speaking, "overdispersed Poisson" is not a single model but rather describes any count-data model for which the variance of the data is ω times the mean, reducing to the Poisson if $\omega = 1$.

A specific model commonly used in this scenario is the so-called negative-binomial distribution:

$$y_i \sim \text{Negative-binomial}\,(\text{mean} = u_i \exp(X_i\beta),\ \text{overdispersion} = \omega).$$

Unfortunately, the negative-binomial distribution is conventionally expressed not based on its mean and overdispersion but rather in terms of parameters a and b, where the mean of the distribution is a/b and the overdispersion is $1 + 1/b$. One must check the parameterization when fitting such models, and it can be helpful to double-check by simulating datasets from the fitted model and checking that they look like the actual data (see Section 8.3).

We return to the police stops example, correcting for overdispersion using a multilevel model, in Section 15.1.

6.3 Logistic-binomial model

Chapter 5 discussed logistic regression for binary (Yes/No or 0/1) data. The logistic model can also be used for count data, using the binomial distribution (see page 16) to model the number of "successess" out of a specified number of possibilities, with the probability of success being fit to a logistic regression.

The binomial model for count data, applied to death sentences

We illustrate binomial logistic regression in the context of a study of the proportion of death penalty verdicts that were overturned, in each of 34 states in the 23 years, 1973–1995. The units of this analysis are the $34 \times 23 = 784$ state-years (actually, we only have $n = 450$ state-years in our analysis, since different states have restarted the death penalty at different times since 1973). For each state-year i, we label n_i as the number of death sentences in that state in that year and $\overline{y_i}$ as the number of these verdicts that were later overturned by higher courts. Our model has the form

$$\begin{aligned} y_i &\sim \text{Binomial}(n_i, p_i) \\ p_i &= \text{logit}^{-1}(X_i\beta), \end{aligned} \tag{6.6}$$

where X is a matrix of predictors. To start, we use

- A constant term
- 33 indicators for states
- A time trend for years (that is, a variable that equals 1 for 1973, 2 for 1974, 3 for 1975, and so on).

This model could also be written as

$$\begin{aligned} y_{st} &\sim \text{Binomial}(n_{st}, p_{st}) \\ p_{st} &= \text{logit}^{-1}(\mu + \alpha_s + \beta t), \end{aligned}$$

with subscripts s for state and t for time (that is, year−1972). We prefer the form (6.6) because of its greater generality. But it is useful to be able to go back and forth between the two formulations.

Overdispersion

When logistic regression is applied to count data, it is possible—in fact, usual—for the data to have more variation than is explained by the model. This overdispersion problem arises because the logistic regression model does not have a variance parameter σ.

More specifically, if data y have a binomial distribution with parameters n and p, then the mean of y is np and the standard deviation of y is $\sqrt{np(1-p)}$. As in model (6.4), we define the standardized residual for each data point i as

$$\begin{aligned} z_i &= \frac{y_i - \hat{y}_i}{\text{sd}(\hat{y}_i)} \\ &= \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}, \end{aligned} \tag{6.7}$$

where $p_i = \text{logit}^{-1}(X_i\hat{\beta})$. If the binomial model is true, then the z_i's should be approximately independent, each with mean 0 and standard deviation 1.

As with the Poisson model, we can then compute the estimated overdispersion $\frac{1}{n-k}\sum_{i=1}^n z_i^2$ (see model (6.5) on page 114) and formally test for overdispersion by comparing $\sum_{i=1}^n z_i^2$ to a χ^2_{n-k} distribution. (The n here represents the number of data points and is unrelated to the notation n_i in models (6.6) and (6.7) referring to the number of cases in state-year i.)

In practice, overdispersion happens almost all the time that logistic regression (or Poisson regression, as discussed in Section 6.2) is applied to count data. In the more general family of distributions known as overdispersed models, the standard deviation can have the form $\sqrt{\omega np(1-p)}$, where $\omega > 1$ is known as the *overdispersion parameter.* The overdispersed model reduces to binomial logistic regression when $\omega = 1$.

Adjusting inferences for overdispersion

As with Poisson regression, a simple correction for overdispersion is to multiply the standard errors of all the coefficient estimates by the square root of the estimated overdispersion (6.5). Without this adjustment, the confidence intervals would be too narrow, and inferences would be overconfident.

Overdispersed binomial regressions can be fit in R using the `glm()` function with the `quasibinomial(link="logit")` family. A corresponding distribution is the beta-binomial.

Binary-data model as a special case of the count-data model

Logistic regression for binary data as in Chapter 5 is a special case of the binomial form (6.6) with $n_i \equiv 1$ for all i. Overdispersion at the level of the individual data points cannot occur in the binary model, which is why we did not introduce overdispersed models in Chapter 5.

Count-data model as a special case of the binary-data model

Conversely, the binomial model (6.6) can be expressed in the binary-data form (5.1) by considering each of the n_i cases as a separate data point. The sample size of this expanded regression is $\sum_i n_i$, and the data points are 0's and 1's: each unit i corresponds to y_i ones and $n_i - y_i$ zeroes. Finally, the X matrix is expanded to have $\sum_i n_i$ rows, where the i^{th} row of the original X matrix becomes n_i identical rows in the expanded matrix. In this parameterization, overdispersion could be included in a multilevel model by creating an index variable for the original measurements (in the death penalty example, taking on the values $1, \ldots, 450$) and including a varying coefficient or error term at this level.

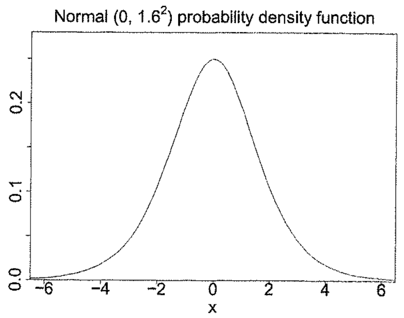

Figure 6.2 *Normal density function with mean 0 and standard deviation 1.6. For most practical purposes, this is indistinguishable from the logistic density (Figure 5.5 on page 85). Thus we can interpret coefficients in probit models as logistic regression coefficients divided by 1.6.*

6.4 Probit regression: normally distributed latent data

The *probit* model is the same as the logit, except it replaces the logistic by the normal distribution (see Figure 5.5). We can write the model directly as

$$\Pr(y_i = 1) = \Phi(X_i\beta),$$

where Φ is the normal cumulative distribution function. In the latent-data formulation,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim \mathrm{N}(0,1), \end{aligned} \qquad (6.8)$$

that is, a normal distribution for the latent errors with mean 0 and standard deviation 1.

More generally, the model can have an error variance, so that the last line of (6.8) is replaced by

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2),$$

but then σ is nonidentified, because the model is unchanged if we multiply σ by some constant c and then multiply the vector β by c also. Hence we need some restriction on the parameters, and the standard approach is to fix $\sigma = 1$ as in (6.8).

Probit or logit?

As is shown in Figure 6.2 (compare to Figure 5.5 on page 85), the probit model is close to the logit with the residual standard deviation of ϵ set to 1 rather than 1.6. As a result, coefficients in a probit regression are typically close to logistic regression coefficients divided by 1.6. For example, here is the probit version of the logistic regression model on page 88 for well switching:

R output

```
glm(formula = switch ~ dist100, family=binomial(link="probit"))
            coef.est coef.se
(Intercept)  0.38     0.04
dist100     -0.39     0.06
  n = 3020, k = 2
  residual deviance = 4076.3, null deviance = 4118.1 (difference = 41.8)
```

For the examples we have seen, the choice of logit or probit model is a matter of taste or convenience, for example, in interpreting the latent normal errors of probit models. When we see probit regression coefficients, we can simply multiply them by 1.6 to obtain the equivalent logistic coefficients. For example, the model we have just fit, $\Pr(y = 1) = \Phi(0.38 - 0.39x)$, is essentially equivalent to the logistic model $\Pr(y = 1) = \text{logit}^{-1}(1.6(0.38 - 0.39x)) = \text{logit}^{-1}(0.61 - 0.62x)$, which indeed is the logit model estimated on page 88.

6.5 Ordered and unordered categorical regression

Logistic and probit regression can be extended to multiple categories, which can be ordered or unordered. Examples of ordered categorical outcomes include Democrat, Independent, Republican; Yes, Maybe, No; Always, Frequently, Often, Rarely, Never. Examples of unordered categorical outcomes include Liberal, Labor, Conservative; Football, Basketball, Baseball, Hockey; Train, Bus, Automobile, Walk; White, Black, Hispanic, Asian, Other. We discuss ordered categories first, including an extended example, and then briefly discuss regression models for unordered categorical variables.

The ordered multinomial logit model

Consider a categorical outcome y that can take on the values $1, 2, \ldots, K$. The ordered logistic model can be written in two equivalent ways. First we express it as a series of logistic regressions:

$$\begin{aligned} \Pr(y > 1) &= \text{logit}^{-1}(X\beta) \\ \Pr(y > 2) &= \text{logit}^{-1}(X\beta - c_2) \\ \Pr(y > 3) &= \text{logit}^{-1}(X\beta - c_3) \\ &\cdots \\ \Pr(y > K{-}1) &= \text{logit}^{-1}(X\beta - c_{K-1}). \end{aligned} \qquad (6.9)$$

The parameters c_k (which are called thresholds or *cutpoints*, for reasons which we shall explain shortly) are constrained to increase: $0 = c_1 < c_2 < \cdots < c_{K-1}$, because the probabilities in (6.9) are strictly decreasing (assuming that all K outcomes have nonzero probabilities of occurring). Since c_1 is defined to be 0, the model with K categories has $K-2$ free parameters c_k in addition to β. This makes sense since $K{=}2$ for the usual logistic regression, for which only β needs to be estimated.

The cutpoints $c_2, \ldots, c_{K-1}$ can be estimated using maximum likelihood, simultaneously with the coefficients β. For some datasets, however, the parameters can be nonidentified, as with logistic regression for binary data (see Section 5.8).

The expressions in (6.9) can be subtracted to get the probabilities of individual outcomes:

$$\begin{aligned} \Pr(y = k) &= \Pr(y > k{-}1) - \Pr(y > k) \\ &= \text{logit}^{-1}(X\beta - c_{k-1}) - \text{logit}^{-1}(X\beta - c_k). \end{aligned}$$

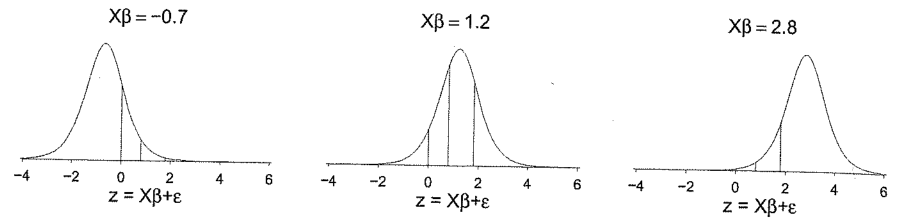

Figure 6.3 *Illustration of cutpoints in an ordered categorical logistic model. In this example, there are $K = 4$ categories and the cutpoints are $c_1 = 0, c_2 = 0.8, c_3 = 1.8$. The three graphs illustrate the distribution of the latent outcome z corresponding to three different values of the linear predictor, $X\beta$. For each, the cutpoints show where the outcome y will equal 1, 2, 3, or 4.*

Latent variable interpretation with cutpoints

The ordered categorical model is easiest to understand by generalizing the latent variable formulation (5.4) to K categories:

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i < 0 \\ 2 & \text{if } z_i \in (0, c_2) \\ 3 & \text{if } z_i \in (c_2, c_3) \\ & \dots \\ K\!-\!1 & \text{if } z_i \in (c_{K-2}, c_{K-1}) \\ K & \text{if } z_i > c_{K-1} \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned} \tag{6.10}$$

with independent errors ϵ_i that have the logistic distribution, as in (5.4).

Figure 6.3 illustrates the latent variable model and shows how the distance between any two adjacent cutpoints c_{k-1}, c_k affects the probability that $y = k$. We can also see that if the linear predictor $X\beta$ is high enough, y will almost certainly take on the highest possible value, and if $X\beta$ is low enough, y will almost certainly equal the lowest possible value.

Example: storable votes

We illustrate ordered categorical data analysis with a study from experimental economics, on the topic of "storable votes." This example is somewhat complicated, and illustrates both the use and potential limitations of the ordered logistic model. In the experiment under study, college students were recruited to play a series of voting games. In each game, a set of k players vote on two issues, with the twist being that each player is given a total of 4 votes. On the first issue, a player has the choice of casting 1, 2, or 3 votes, with the remaining votes cast on the second issue. The winning side of each issue is decided by majority vote, at which point the players on the winning side each get positive payoffs, which are drawn from a uniform distribution on the interval [1, 100].

To increase their expected payoffs, players should follow a strategy of casting more votes for issues where their potential payoffs are higher. The way this experiment is conducted, the players are told the distribution of possible payoffs, and they are told their potential payoff for each issue just before the vote. Thus, in making the choice of how many votes to cast in the first issue, each player knows his or her potential payoff for that vote only. Then, the players are told their potential payoffs for the second vote, but no choice is involved at this point since they will automatically spend all their remaining votes. Players' strategies can thus be summarized as their choices of initial votes, $y = 1$, 2, or 3, given their potential payoff, x.

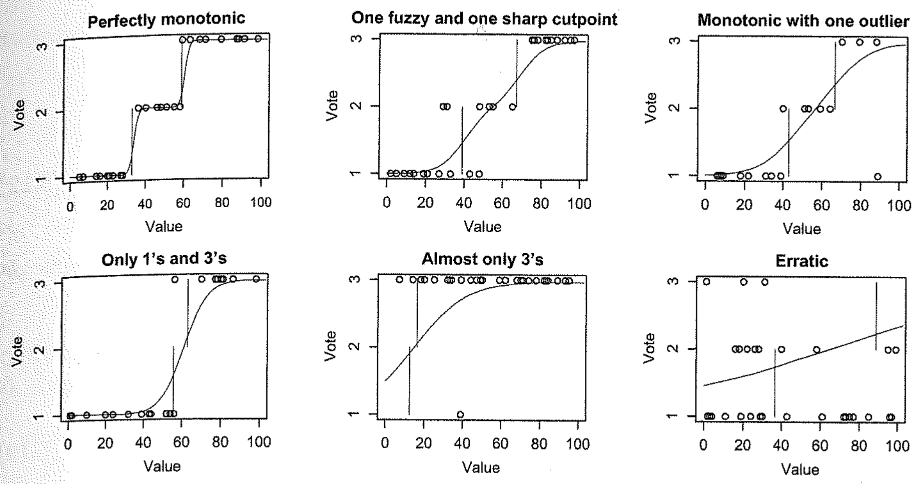

Figure 6.4 *Data from some example individuals in the storable votes study. Vertical lines show estimated cutpoints, and curves show expected responses as estimated using ordered logistic regressions. The two left graphs show data that fit the model reasonably well; the others fit the model in some ways but not perfectly.*

Figure 6.4 graphs the responses from six of the hundred or so students in the experiment, with these six chosen to represent several different patterns of data. We were not surprised to see that responses were generally monotonic—that is, students tend to spend more votes when their potential payoff is higher—but it was interesting to see the variety of approximately monotonic strategies that were chosen.

As is apparent in Figure 6.4, most individuals' behaviors can be summarized by three parameters—the cutpoint between votes of 1 and 2, the cutpoint between 2 and 3, and the fuzziness of these divisions. The two cutpoints characterize the chosen monotone strategy, and the sharpness of the divisions indicates the consistency with which the strategy is followed.

Three parameterizations of the ordered logistic model. It is convenient to model the responses using an ordered logit, using a parameterization slightly different from that of model (6.10) to match up with our understanding of the monotone strategies. The model is

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i < c_{1.5} \\ 2 & \text{if } z_i \in (c_{1.5}, c_{2.5}) \\ 3 & \text{if } z_i > c_{2.5} \end{cases} \\ z_i &\sim \text{logistic}(x_i, \sigma^2). \end{aligned} \tag{6.11}$$

In this model, the cutpoints $c_{1.5}$ and $c_{2.5}$ are on the 1–100 scale of the data x, and the scale σ of the errors ϵ corresponds to the fuzziness of the cutpoints.

This model has the same number of parameters as the conventional parameterization (6.10)—two regression coefficients have disappeared, while one additional free cutpoint and an error variance have been added. Here is model (6.10) with

$K = 3$ categories and one predictor x,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i < 0 \\ 2 & \text{if } z_i \in (0, c_2) \\ 3 & \text{if } z_i > c_2 \end{cases} \\ z_i &= \alpha + \beta x + \epsilon_i, \end{aligned} \tag{6.12}$$

with independent errors $\epsilon_i \sim \text{logistic}(0, 1)$.

Yet another version of the model keeps the two distinct cutpoints but removes the constant term, α; thus,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i < c_{1|2} \\ 2 & \text{if } z_i \in (0, c_{2|3}) \\ 3 & \text{if } z_i > c_{2|3} \end{cases} \\ z_i &= \beta x + \epsilon_i, \end{aligned} \tag{6.13}$$

with independent errors $\epsilon_i \sim \text{logistic}(0, 1)$.

The three models are in fact equivalent, with z_i/β in (6.13) and $(z_i - \alpha)/\beta$ in (6.12) corresponding to z_i in (6.11) and the parameters matching up as follows:

Model (6.11)	Model (6.12)	Model (6.13)
$c_{1.5}$	$-\alpha/\beta$	$-c_{1\|2}/\beta$
$c_{2.5}$	$(c_2 - \alpha)/\beta$	$-c_{2\|3}/\beta$
σ	$1/\beta$	$1/\beta$

We prefer parameterization (6.11) because we can directly interpret $c_{1.5}$ and $c_{2.5}$ as thresholds on the scale of the input x, and σ corresponds to the gradualness of the transitions from 1's to 2's and from 2's to 3's. It is sometimes convenient, however, to fit the model using the standard parameterizations (6.12) and (6.13), and so it is helpful to be able to go back and forth between the models.

Fitting the model in R. We can fit ordered logit (or probit) models using the `polr` ("proportional odds logistic regression") function, which is part of the `MASS` package in R. We illustrate with data from one of the persons in the storable votes study:

R code

```
polr (factor(y) ~ x)
```

which yields

R output

```
Coefficients:
         x
0.07911799

Intercepts:
     1|2      2|3
1.956285 4.049963
```

From the output we can see this has fitted a model of the form (6.13), with estimates $\hat{\beta} = 0.079$, $\hat{c}_{1|2} = 1.96$ and $\hat{c}_{2|3} = 4.05$. Transforming to model (6.11) using the table of the three models, we get $\hat{c}_{1.5} = 1.96/0.079 = 24.8$, $\hat{c}_{2.5} = 4.03/0.079 = 51.3$, and $\hat{\sigma} = 1/0.079 = 12.7$.

Displaying the fitted model. Figure 6.4 shows the cutpoints $c_{1.5}, c_{2.5}$ and expected votes $\text{E}(y)$ as a function of x, as estimated from the data from each of several students. From the model (6.11), the expected votes can be written as

$$\begin{aligned} \text{E}(y|x) &= 1 \cdot \Pr(y = 1|x) + 2 \cdot \Pr(y = 2|x) + 3 \cdot \Pr(y = 3|x) \\ &= 1 \cdot \left(1 - \text{logit}^{-1}\left(\frac{x - c_{1.5}}{\sigma}\right)\right) + \\ &\quad + 2 \cdot \left(\text{logit}^{-1}\left(\frac{x - c_{1.5}}{\sigma}\right) - \text{logit}^{-1}\left(\frac{x - c_{2.5}}{\sigma}\right)\right) + \\ &\quad + 3 \cdot \text{logit}^{-1}\left(\frac{x - c_{2.5}}{\sigma}\right), \end{aligned} \tag{6.14}$$

where $\text{logit}^{-1}(x) = e^x/(1+e^x)$ is the logistic curve displayed in Figure 5.2a on page 80. Expression (6.14) looks complicated but is easy to program as a function in R:

R code

```
expected <- function (x, c1.5, c2.5, sigma){
  p1.5 <- invlogit ((x-c1.5)/sigma)
  p2.5 <- invlogit ((x-c2.5)/sigma)
  return ((1*(1-p1.5) + 2*(p1.5-p2.5) + 3*p2.5))
}
```

The data, cutpoints, and curves in Figure 6.4 can then be plotted as follows:

R code

```
plot (x, y, xlim=c(0,100), ylim=c(1,3), xlab="Value", ylab="Vote")
lines (rep (c1.5, 2), c(1,2))
lines (rep (c2.5, 2), c(2,3))
curve (expected (x, c1.5, c2.5, sigma), add=TRUE)
```

Having displayed these estimates for individuals, the next step is to study the distribution of the parameters in the population, to understand the range of strategies applied by the students. In this context, the data have a multilevel structure—30 observations for each of several students—and we pursue this example further in Section 15.2 in the chapter on multilevel generalized linear models.

Alternative approaches to modeling ordered categorical data

Ordered categorical data can be modeled in several ways, including:

- Ordered logit model with $K-1$ cutpoint parameters, as we have just illustrated.
- The same model in probit form.
- Simple linear regression (possibly preceded by a simple transformation of the outcome values). This can be a good idea if the number of categories is large and if they can be considered equally spaced. This presupposes that a reasonable range of the categories is actually used. For example, if ratings are on a 1 to 10 scale, but in practice always equal 9 or 10, then a linear model probably will not work well.
- Separate logistic regressions—that is, a logistic regression model for $y = 1$ versus $y = 2, \ldots, K$; then, if $y \geq 2$, a logistic regression for $y = 2$ versus $y = 3, \ldots, K$; and so on up to a model, if $y \geq K - 1$ for $y = K-1$ versus $y = K$. Or this can be set up using the probit model. Separate logistic (or probit) regressions have the advantage of more flexibility in fitting data but the disadvantage of losing the simple latent-variable interpretation of the cutpoint model we have described.
- Finally, robit regression, which we discuss in Section 6.6, is a competitor to logistic regression that accounts for occasional aberrant data such as the outlier in the upper-right plot of Figure 6.4.

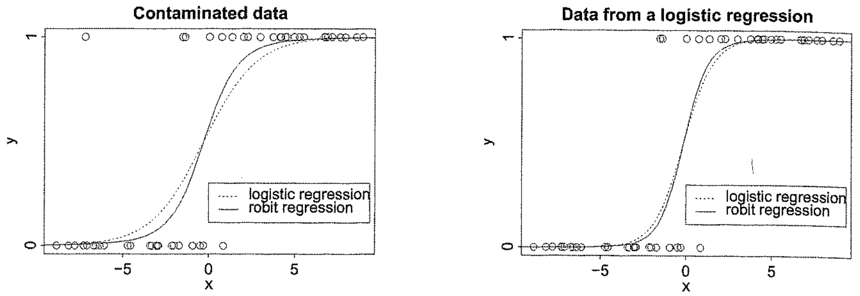

Figure 6.5 *Hypothetical data to be fitted using logistic regression: (a) a dataset with an "outlier" (the unexpected $y = 1$ value near the upper left); (b) data simulated from a logistic regression model, with no outliers. In each plot, the dotted and solid lines show the fitted logit and robit regressions, respectively. In each case, the robit line is steeper—especially for the contaminated data—because it effectively downweights the influence of points that do not appear to fit the model.*

Unordered categorical regression

As discussed at the beginning of Section 6.5, it is sometimes appropriate to model discrete outcomes as unordered. An example that arose in our research was the well-switching problem. As described in Section 5.4, households with unsafe wells had the option to switch to safer wells. But the actual alternatives are more complicated and can be summarized as: (0) do nothing, (1) switch to an existing private well, (2) switch to an existing community well, (3) install a new well yourself. If these are coded as 0, 1, 2, 3, then we can model $\Pr(y \geq 1), \Pr(y \geq 2|y \geq 1), \Pr(y = 3|y \geq 2)$. Although the four options could be considered to be ordered in some way, it does not make sense to apply the ordered multinomial logit or probit model, since different factors likely influence the three different decisions. Rather, it makes more sense to fit separate logit (or probit) models to each of the three components of the decision: (a) do you switch or do nothing? (b) if you switch, do you switch to an existing well or build a new well yourself? (c) if you switch to an existing well, is it a private or community well? More about this important category of model can be found in the references at the end of this chapter.

6.6 Robust regression using the t model

The t distribution instead of the normal

When a regression model can have occasional very large errors, it is generally more appropriate to use a Student-t rather than normal distribution for the errors. The basic form of the regression is unchanged—$y = X\beta + \epsilon$—but with a different distribution for the ϵ's and thus a slightly different method for estimating β (see the discussion of maximum likelihood estimation in Chapter 18) and a different distribution for predictions. Regressions estimated using the t model are said to be *robust* in that the coefficient estimates are less influenced by individual outlying data points. Regressions with t errors can be fit using the `tlm()` function in the `hett` package in R.

Robit instead of logit or probit

Logistic regression (and the essentially equivalent probit regression) are flexible and convenient for modeling binary data, but they can run into problems with outliers. Outliers are usually thought of as extreme observations, but in the context of discrete data, an "outlier" is more of an *unexpected* observation. Figure 6.5a illustrates, with data simulated from a logistic regression, with an extreme point switched from 0 to 1. In the context of the logistic model, an observation of $y = 1$ for this value of x would be extremely unlikely, but in real data this sort of "misclassification" can definitely occur. Hence this graph represents the sort of data to which we might fit a logistic regression, even though this model is not exactly appropriate.

For another illustration of a logistic regression with an aberrant data point, see the upper-right plot in Figure 6.4. That is an example with three outcomes; for simplicity, we restrict our attention here to binary outcomes.

Logistic regression can be conveniently "robustified" by generalizing the latent-data formulation (5.4):

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned}$$

to give the latent errors ϵ a t distribution:

$$\epsilon_i \sim t_\nu\left(0, \frac{\nu - 2}{\nu}\right), \tag{6.15}$$

with the degrees-of-freedom parameter $\nu > 2$ estimated from the data and the t distribution scaled so that its standard deviation equals 1.

The t model for the ϵ_i's allows the occasional unexpected prediction—a positive value of z for a highly negative value of the linear predictor $X\beta$, or vice versa. Figure 6.5a illustrates with the simulated "contaminated" dataset: the solid line shows $\Pr(y = 1)$ as a function of the x for the fitted robit regression, and it is quite a bit steeper than the fitted logistic model. The t distribution effectively downweights the discordant data point so that the model better fits the main part of the data.

Figure 6.5b shows what happens with data that actually come from a logistic model: here, the robit model is close to the logit, which makes sense since it does not find discrepancies.

Mathematically, the robit model can be considered as a generalization of probit and an approximate generalization of logit. Probit corresponds to the degrees of freedom $\nu = \infty$, and logit is very close to the robit model with $\nu = 7$.

6.7 Building more complex generalized linear models

The models we have considered so far can handle many regression problems in practice. For continuous data we start with linear regression with normal errors, consider appropriate transformations and interactions as discussed in Chapter 4, and switch to a t error model for data with occasional large errors. For binary data we use logit, probit, or perhaps robit, again transforming input variables and considering residual plots as discussed in Chapter 5. For count data, the starting points are the overdispersed binomial and Poisson distributions, and for discrete outcomes with more than two categories we can fit ordered or unordered multinomial logit or probit regression. Here we briefly describe some situations where it is helpful to consider other models.

Mixed discrete/continuous data

Earnings is an example of an outcome variable with both discrete and continuous aspects. In our earnings and height regressions in Chapter 4, we preprocessed the data by removing all respondents with zero earnings. In general, however, it can be appropriate to model a variable such as earnings in two steps: first a logistic regression for $\Pr(y > 0)$, then a linear regression on $\log(y)$, conditional on $y > 0$. Predictions for such a model then must be done in two steps, most conveniently using simulation (see Chapter 7).

When modeling an outcome in several steps, programming effort is sometimes required to convert inferences on to the original scale of the data. For example, in a two-step model for predicting earnings given height and sex, we first use a logistic regression to predict whether earnings are positive:

R code

```
earn.pos <- ifelse (earnings>0, 1, 0)
fit.1a <- glm (earn.pos ~ height + male, family=binomial(link="logit"))
```

yielding the fit

R output

```
              coef.est coef.se
(Intercept)    -3.85    2.07
height          0.08    0.03
male            1.70    0.32
  n = 1374, k = 3
  residual deviance = 988.3, null deviance = 1093.2 (difference = 104.9)
```

We then fit a linear regression to the logarithms of positive earnings:

R code

```
log.earn <- log(earnings)
fit.1b <- lm (log.earn ~ height + male, subset = earnings>0)
```

yielding the fit

R output

```
              coef.est coef.se
(Intercept)     8.12    0.60
height          0.02    0.01
male            0.42    0.07
  n = 1187, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

Thus, for example, a 66-inch-tall woman has a probability $\text{logit}^{-1}(-3.85 + 0.08 \cdot 66 + 1.70 \cdot 0) = 0.81$, or an 81% chance, of having positive earnings. If her earnings are positive, their predicted value is $\exp(8.12+0.02{\cdot}66+0.42{\cdot}0) = 12600$. Combining these gives a mixture of a spike at 0 and a lognormal distribution, which is most easily manipulated using simulations, as we discuss in Sections 7.4 and 25.4.

Latent-data models. Another way to model mixed data is through latent data, for example positing an "underlying" income level z_i—the income that person i would have if he or she were employed—that is observed only if $y_i > 0$. *Tobit regression* is one such model that is popular in econometrics.

Cockroaches and the zero-inflated Poisson model

The binomial and Poisson models, and their overdispersed generalizations, all can be expressed in terms of an underlying continuous probability or rate of occurrence of an event. Sometimes, however, the underlying rate itself has discrete aspects.

For example, in a study of cockroach infestation in city apartments, each apartment i was set up with traps for several days. We label u_i as the number of trap-days and y_i as the number of cockroaches trapped. With a goal of predicting cockroach infestation given predictors X (including income and ethnicity of the apartment dwellers, indicators for neighborhood, and measures of quality of the apartment), we would start with the model

$$y_i \sim \text{overdispersed Poisson}\,(u_i e^{X_i\beta},\, \omega). \tag{6.16}$$

It is possible, however, for the data to have more zeroes (that is, apartments i with cockroach counts $y_i = 0$) than predicted by this model.[3] A natural explanation is that some apartments have truly a zero (or very near-zero) rate of cockroaches, whereas others simply have zero counts from the discreteness of the data. The *zero-inflated* model places (6.16) into a mixture model:

$$y_i \begin{cases} = 0, & \text{if } S_i = 0 \\ \sim \text{overdispersed Poisson}\,(u_i e^{X_i\beta},\, \omega), & \text{if } S_i = 1. \end{cases}$$

Here, S_i is an indicator of whether apartment i has any cockroaches at all, and it could be modeled using logistic regression:

$$\Pr(S_i = 1) = \text{logit}^{-1}(X_i\gamma),$$

where γ is a new set of regression coefficients for this part of the model. Estimating this two-stage model is not simple—the S_i's are not observed and so one cannot directly estimate γ; and we do not know which zero observations correspond to $S_i = 0$ and which correspond to outcomes of the Poisson distribution, so we cannot directly estimate β. Some R functions have been written to fit such models and they can also be fit using Bugs.

Other models

The basic choices of linear, logistic, and Poisson models, along with mixtures of these models and their overdispersed, robust, and multinomial generalizations, can handle many regression problems. However, other distributional forms have been used for specific sorts of data; these include exponential, gamma, and Weibull models for waiting-time data, and hazard models for survival data. More generally, nonparametric models including generalized additive models, neural networks, and many others have been developed for going beyond the generalized linear modeling framework by allowing data-fitted nonlinear relations between inputs and the data.

6.8 Constructive choice models

So far we have considered regression modeling as a descriptive tool for studying how an outcome can be predicted given some input variables. A completely different approach, sometimes applicable to choice data such as in the examples in Chapters 5 and 6 on logistic regression and generalized linear models, is to model the decisions as a balancing of goals or utilities.

We demonstrate this idea using the example of well switching in Bangladesh (see Section 5.4). How can we understand the relation between distance, arsenic level, and the decision to switch? It makes sense that people with higher arsenic levels would be more likely to switch, but what coefficient values should we expect? Should the relation be on the log or linear scale? The actual health risk is believed

[3] In our actual example, the overdispersed Poisson model did a reasonable job predicting the number of zeroes; see page 161. But in other similar datasets the zero-inflated model can both make sense and fit data well, hence our presentation here.

to be linear in arsenic concentration; does that mean that a logarithmic model is inappropriate? Such questions can be addressed using a model for individual decisions.

To set up a *choice model*, we must specify a *value function*, which represents the strength of preference for one decision over the other—in this case, the preference for switching as compared to not switching. The value function is scaled so that zero represents indifference, positive values correspond to a preference for switching, and negative values result in not switching. This model is thus similar to the latent-data interpretation of logistic regression (see page 85); and in fact that model is a special case, as we shall see here.

Logistic or probit regression as a choice model in one dimension

There are simple one-dimensional choice models that reduce to probit or logit regression with a single predictor, as we illustrate with the model of switching given distance to nearest well. From page 88, the logistic regression is

R output

```
glm(formula = switch ~ dist100, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)    0.61    0.06
dist100       -0.62    0.10
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

Now let us think about it from first principles as a decision problem. For household i, define

- a_i = the benefit of switching from an unsafe to a safe well
- $b_i + c_i x_i$ = the cost of switching to a new well a distance x_i away.

We are assuming a utility theory in which the benefit (in reduced risk of disease) can be expressed on the same scale as the cost (the inconvenience of no longer using one's own well, plus the additional effort—proportional to distance—required to carry the water).

Logit model. Under the utility model, household i will switch if $a_i > b_i + c_i x_i$. However, we do not have direct measurements of the a_i's, b_i's, and c_i's. All we can learn from the data is the probability of switching as a function of x_i; that is,

$$\Pr(\text{switch}) = \Pr(y_i = 1) = \Pr(a_i > b_i + c_i x_i), \qquad (6.17)$$

treating a_i, b_i, c_i as random variables whose distribution is determined by the (unknown) values of these parameters in the population.

Expression (6.17) can be written as

$$\Pr(y_i = 1) = \Pr\left(\frac{a_i - b_i}{c_i} > x_i\right),$$

a re-expression that is useful in that it puts all the random variables in the same place and reveals that the population relation between y and x depends on the distribution of $(a - b)/c$ in the population.

For convenience, label $d_i = (a_i - b_i)/c_i$: the net benefit of switching to a neighboring well, divided by the cost per distance traveled to a new well. If d_i has a logistic distribution in the population, and if d is independent of x, then $\Pr(y = 1)$ will have the form of a logistic regression on x, as we shall show here.

If d_i has a logistic distribution with center μ and scale σ, then $d_i = \mu + \sigma\epsilon_i$, where ϵ_i has the unit logistic density; see Figure 5.2 on page 80. Then

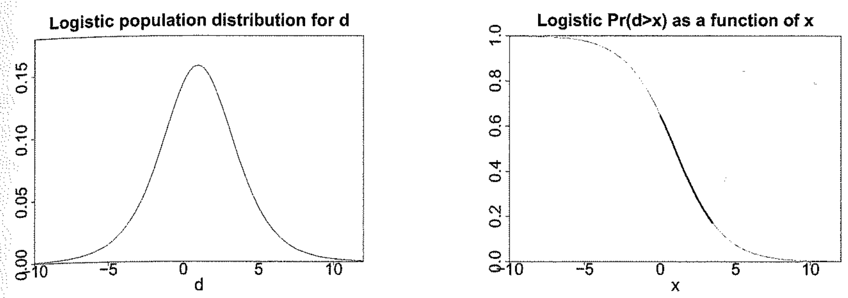

Figure 6.6 *(a) Hypothesized logistic distribution of $d_i = (a_i - b_i)/c_i$ in the population and (b) corresponding logistic regression curve of the probability of switching given distance. These both correspond to the model, $\Pr(y_i = 1) = \Pr(d_i > x_i) = logit^{-1}(0.61 - 0.62x)$. The dark part of the curve in (b) corresponds to the range of x (distance in 100-meter units) in the well-switching data; see Figure 5.9 on page 89.*

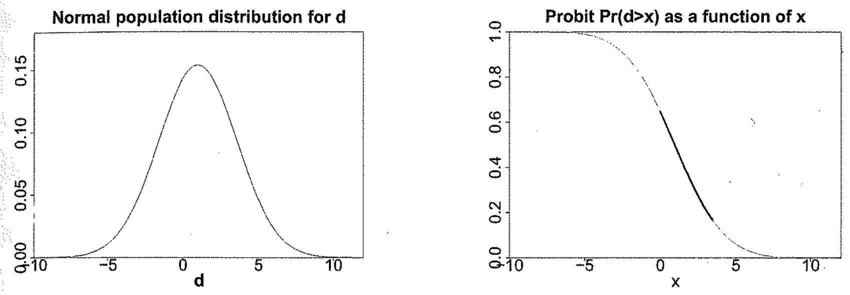

Figure 6.7 *(a) Hypothesized normal distribution of $d_i = (a_i - b_i)/c_i$ with mean 0.98 and standard deviation 2.6 and (b) corresponding probit regression curve of the probability of switching given distance. These both correspond to the model, $\Pr(y_i = 1) = \Pr(d_i > x_i) = \Phi(0.38 - 0.39x)$. Compare to Figure 6.6.*

$$\begin{aligned}\Pr(\text{switch}) = \Pr(d_i > x) &= \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right)\\ &= \text{logit}^{-1}\left(\frac{\mu - x}{\sigma}\right) = \text{logit}^{-1}\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}x\right),\end{aligned}$$

which is simply a logistic regression with coefficients μ/σ and $-1/\sigma$. We can then fit the logistic regression and solve for μ and σ. For example, the well-switching model, $\Pr(y = 1) = \text{logit}^{-1}(0.61 - 0.62x)$, corresponds to $\mu/\sigma = 0.61$ and $-1/\sigma = -0.62$; thus $\sigma = 1/0.62 = 1.6$ and $\mu = 0.61/0.62 = 0.98$. Figure 6.6 shows the distribution of d, along with the curve of $\Pr(d > x)$ as a function of x.

Probit model. A similar model is obtained by starting with a normal distribution for the utility parameter: $d \sim \text{N}(\mu, \sigma^2)$. In this case,

$$\begin{aligned}\Pr(\text{switch}) = \Pr(d_i > x) &= \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right)\\ &= \Phi\left(\frac{\mu - x}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}x\right),\end{aligned}$$

which is simply a probit regression. The model $\Pr(y = 1) = \Phi(0.38 - 0.39x)$ corresponds to $\mu/\sigma = 0.38$ and $-1/\sigma = -0.39$; thus $\sigma = 1/0.39 = 2.6$ and

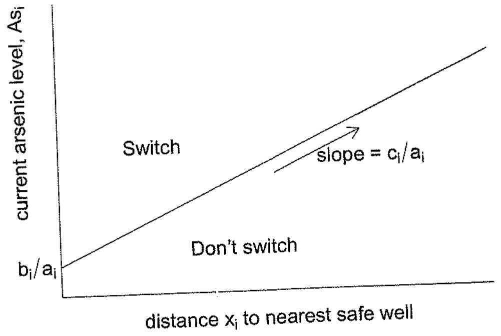
current arsenic level, As$_i$
Switch
slope = c$_i$/a$_i$
Don't switch
b$_i$/a$_i$
distance x$_i$ to nearest safe well

Figure 6.8 *Decision options for well switching given arsenic level of current well and distance to the nearest safe well, based on the decision rule: switch if* $a_i \cdot (\mathrm{As})_i > b_i + cx_i$.

$\mu = 0.38/0.39 = 0.98$. Figure 6.7 shows this model, which is nearly identical to the logistic model shown in Figure 6.6.

Choice models, discrete data regressions, and latent data

Logistic regression and generalized linear models are usually set up as methods for estimating the probabilities of different outcomes y given predictors x. A fitted model represents an entire population, with the "error" in the model coming in through probabilities that are not simply 0 or 1 (hence, the gap between data points and fitted curves in graphs such as Figure 5.9 on page 89).

In contrast, choice models are defined at the level of the individual, as we can see in the well-switching example, where each household i has, along with its own data X_i, y_i, its own parameters a_i, b_i, c_i that determine its utility function and thus its decision of whether to switch.

Logistic or probit regression as a choice model in multiple dimensions

We can extend the well-switching model to multiple dimensions by considering the arsenic level of the current well as a factor in the decision.

- $a_i \cdot (\mathrm{As})_i$ = the benefit of switching from an unsafe well with arsenic level As$_i$ to a safe well. (It makes sense for the benefit to be proportional to the current arsenic level, because risk is believed to be essentially proportional to cumulative exposure to arsenic.)
- $b_i + c_i x_i$ = the cost of switching to a new well a distance x_i away.

Household i should then switch if $a_i \cdot (\mathrm{As})_i > b_i + cx_i$—the decision thus depends on the household's arsenic level $(\mathrm{As})_i$, its distance x_i to the nearest well, and its utility parameters a_i, b_i, c_i.

Figure 6.8 shows the decision space for an individual household, depending on its arsenic level and distance to the nearest safe well. Given a_i, b_i, c_i, the decision under this model is deterministic. However, a_i, b_i, c_i are not directly observable—all we see are the decisions ($y_i = 0$ or 1) for households, given their arsenic levels As$_i$ and distances x_i to the nearest safe well.

Certain distributions of (a, b, c) in the population reduce to the fitted logistic regression, for example, if a_i and c_i are constants and b_i/a_i has a logistic distribution that is independent of $(\mathrm{As})_i$ and x_i. More generally, choice models reduce to logistic regressions if the factors come in additively, with coefficients that do not vary in the population, and if there is a fixed cost (b_i in this example) that has a logistic distribution in the population.

Other distributions of (a, b, c) are possible. The corresponding models can be fit, treating these utility parameters as latent data. There is no easy way of fitting such models using `glm()` in R (except for the special cases that reduce to logit and probit), but they can be fit in Bugs (see Exercise 17.7).

Insights from decision models

A choice model can give us some insight even if we do not formally fit it. For example, in fitting logistic regressions, we found that distance worked well as a linear predictor, whereas arsenic level fit better on the logarithmic scale. A simple utility analysis would suggest that both these factors should come in linearly, and the transformation for arsenic suggests that people are (incorrectly) perceiving the risks on a logarithmic scale—seeing the difference between 4 to 8, say, as no worse than the difference between 1 and 2. (In addition, our residual plot showed the complication that people seem to underestimate risks from arsenic levels very close to 0.5. And behind this is the simplifying assumption that all wells with arsenic levels below 0.5 are "safe.")

We can also use the utility model to interpret the coefficient for education in the model—more educated people are more likely to switch, indicating that their costs of switching are lower, or their perceived benefits from reducing arsenic exposure are higher. Interactions correspond to dependence among the latent utility parameters in the population.

The model could also be elaborated to consider the full range of individual options, which include doing nothing, switching to an existing private well, switching to an existing community well, or digging a new private well. The decision depends on the cost of walking, perception of health risks, financial resources, and future plans.

6.9 Bibliographic note

The concept of the generalized linear model was introduced by Nelder and Wedderburn (1972) and developed further, with many examples, by McCullagh and Nelder (1989). Dobson (2001) is an accessible introductory text. For more on overdispersion, see Anderson (1988) and Liang and McCullagh (1993). Fienberg (1977) and Agresti (2002) are other useful references.

The death penalty example comes from Gelman, Liebman, et al. (2004). Models for traffic accidents are discussed by Chapman (1973) and Hauer, Ng, and Lovell (1988). For more on the New York City police example, see Spitzer (1999) and Gelman, Fagan, and Kiss (2005).

Maddala (1983) presents discrete-data regressions and choice models from an econometric perspective, and McCullagh (1980) considers general forms for latent-parameter models for ordered data. Amemiya (1981) discusses the factor of 1.6 for converting from logit to probit coefficients.

Walker and Duncan (1967) introduce the ordered logistic regression model, and Imai and van Dyk (2003) discuss the models underlying multinomial logit and probit regression. The storable votes example comes from Casella, Gelman, and Palfrey (2006). See Agresti (2002) and Imai and van Dyk (2003) for more on categorical regression models, ordered and unordered.

Robust regression using the t distribution is discussed by Zellner (1976) and