5
Beyond Straight Lines

Key terms and concepts: logistic regression; link function; generalized linear model

Binary responses

In many situations, we would like to forecast the outcome of a binary event, given some relevant information:

- Given the pattern of word usage and punctuation in an e-mail, is it likely to be spam?
- Given the temperature, pressure, and cloud cover on Christmas Eve, is it likely to snow on Christmas Day?
- Given a person's credit history and income, is he or she likely to default on a mortgage loan?

In all of these cases, the Y variable is the answer to a yes-or-no question. This is a bit different to the kinds of problems we've become used to seeing, where the response is a real number.

Nonetheless, we can still use regression for these problems. Let's suppose, for simplicity's sake, that we have only one predictor x, and that we let $y_i = 1$ for a "yes" and $y_i = 0$ for a "no." One naïve way of forecasting y is simply to plunge ahead with the basic, one-variable regression equation:

$$\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

Since y_i can only take the values 0 or 1, the expected value of y_i is simply a weighted average of these two cases:

$$\begin{aligned} \mathrm{E}(y_i \mid x_i) &= 1 \cdot P(y_i = 1 \mid x_i) + 0 \cdot P(y_i = 0 \mid x_i) \\ &= P(y_i = 1 \mid x_i) \end{aligned}$$

Therefore, the regression equation is just a linear model for the conditional probability that $y_i = 1$, given the predictor x_i:

$$P(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

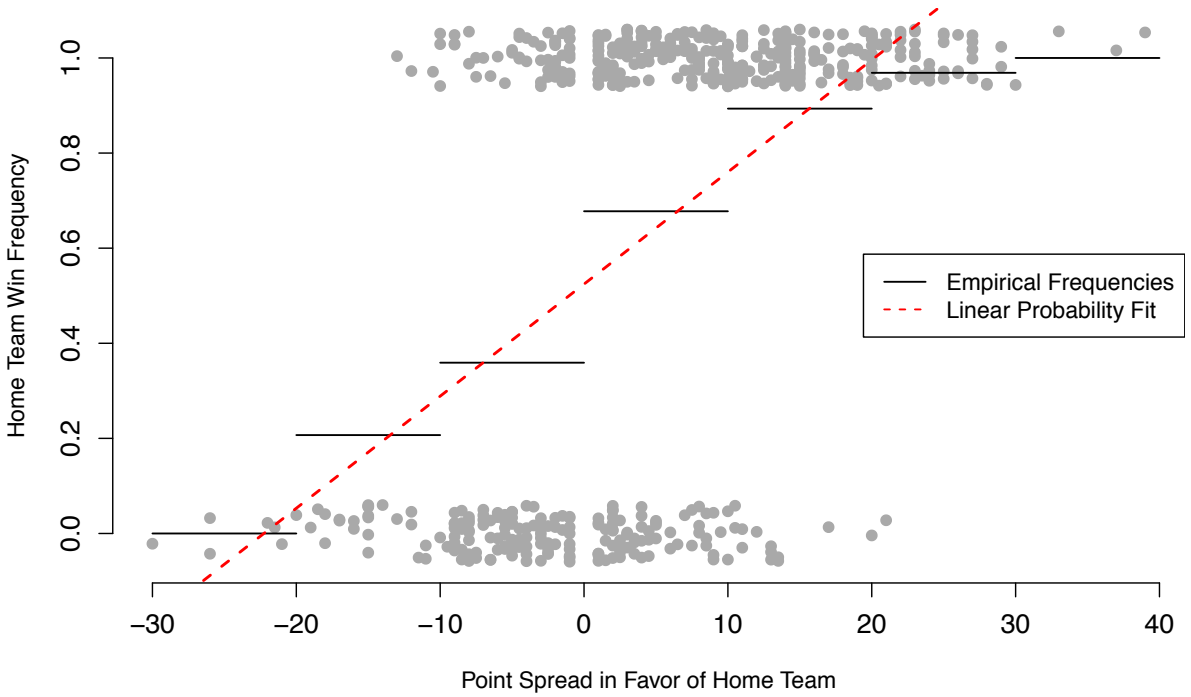

Figure 5.1: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

This model allows us to plug in some value of x_i and read off the forecasted probability of a "yes" answer to whatever yes-or-no question is being posed. It is often called the linear probability model, since the probability of a "yes" varies linearly with x.

Let's try fitting it to some example data to understand how this kind of model behaves. In Table 5.1 on page 131, we see an excerpt of a data set on 553 men's college-basketball games. Our y variable is whether the home team won ($y_i = 1$) or lost ($y_i = 0$). Our x variable is the Las Vegas "point spread" in favor of the home team. The spread indicates the betting market's collective opinion about the home team's expected margin of victory—or defeat, if the spread is negative. Large spreads indicate that one team is heavily favored to win. It is therefore natural to use the Vegas spread to predict the probability of a home-team victory in any particular game.

Figure 5.1 shows each of the 553 results in the data set. The home-team point spread is plotted on the x-axis, while the result of the game is plotted on the y-axis. A home-team win is plotted as a 1, and a loss as a 0. A bit of artificial vertical jitter has been added to the 1's and 0's, just so you can distinguish the individual dots.

The horizontal black lines indicate empirical win frequencies for point spreads in the given range. For example, home teams won about 65% of the time when they were favored by more than 0 points, but less than 10. Similarly, when home teams were 10–20 point underdogs, they won only about 20% of the time.

Finally, the dotted red line is the linear probability fit:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.524435   0.019040   27.54   <2e-16 ***
spread      0.023566   0.001577   14.94   <2e-16 ***
---
Residual standard error: 0.4038 on 551 degrees of freedom
Multiple R-squared: 0.2884
```

This is the result of having regressed the binary y_i's on the point spreads, simply treating the 1's and 0's as if they were real numbers. Under this model, our estimated regression equation is

$$\mathrm{E}(y_i \mid x_i) = P(y_i = 1 \mid x_i) = 0.524 + 0.024 \cdot x_i \,.$$

Plug in an x, and read off the probability of a home-team victory. Here, we would expect the intercept to be 0.5, meaning that the home team should win exactly 50% of the time when the point spread is 0. Of course, because of sampling variability, the estimated intercept $\widehat{\beta}_0$ isn't exactly 0.5. But it's certainly close—about 1 standard error away.

The linear probability model, however, has a serious flaw. Try plugging in $x_i = 21$ and see what happens:

$$P(y_i = 1 \mid x_i = 21) = 0.524 + 0.024 \cdot 21 = 1.028 \,.$$

We get a probability larger than 1, which is clearly nonsensical. We could also get a probability less than zero by plugging in $x_1 = -23$:

$$P(y_i = 1 \mid x_i = -23) = 0.524 - 0.024 \cdot 23 = -.028 \,.$$

Game	Win	Spread
1	0	-7
2	1	7
3	1	17
4	0	9
5	1	-2.5
6	0	-9
7	1	10
8	1	18
9	1	-7.5
10	0	-8
⋮		
552	1	-4.5
553	1	-3

Table 5.1: An excerpt from a data set on 553 NCAA basketball games. "Win" is coded 1 if the home team won the game, and 0 otherwise. "Spread" is the Las Vegas point spread in favor of the home team (at tipoff). Negative point spreads indicate where the visiting team was favored.

The problem is that the straight-line fit does not respect the rule that probabilities must be numbers between 0 and 1. For many values of x_i, it gives results that aren't even mathematically legal.

Link functions and generalized linear models

THE PROBLEM can be summarized as follows. The right-hand side of the regression equation, $\beta_0 + \beta_1 x_i$, can be any real number between $-\infty$ and ∞. But the left-hand side, $P(y_i = 1 \mid x_i)$, must be between 0 and 1. Therefore, we need some transformation g that takes an unconstrained number from the right-hand side, and maps it to a constrained number on the left-hand side:

$$P(y_i \mid x_i) = g(\beta_0 + \beta_1 x_i) \,.$$

Such a function g is called a *link function*; a model that incorporates such a link function is called a *generalized linear model*; and the part inside the parentheses $(\beta_0 + \beta_1 x_i)$ is called the *linear predictor*.

We use link functions and generalized linear models in most situations where we are trying to predict a number that is, for whatever reason, constrained. Here, we're dealing with probabilities, which are constrained to be no smaller than 0 and no larger than 1. Therefore, the function g must map real numbers on $(-\infty, \infty)$ to numbers on $(0, 1)$. It must therefore be shaped a bit like a flattened letter "S," approaching zero for large negative values of the linear predictor, and approaching 1 for large positive values.

Figure 5.2 contains the most common example of such a link function. This is called the *logistic* link, which gives rise to the *logistic regression model*:

$$P(y_i = 1 \mid x_i) = g(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \,.$$

Think of this as just one more transformation, like the logarithm or powers of some predictor x. The only difference is that, in this case, the transformation gets applied to the whole linear predictor at once.

With a little bit of algebra, it is also possible to isolate the linear predictor $\beta_0 + \beta_1 x_i$ on one side of the equation. If we let p_i denote

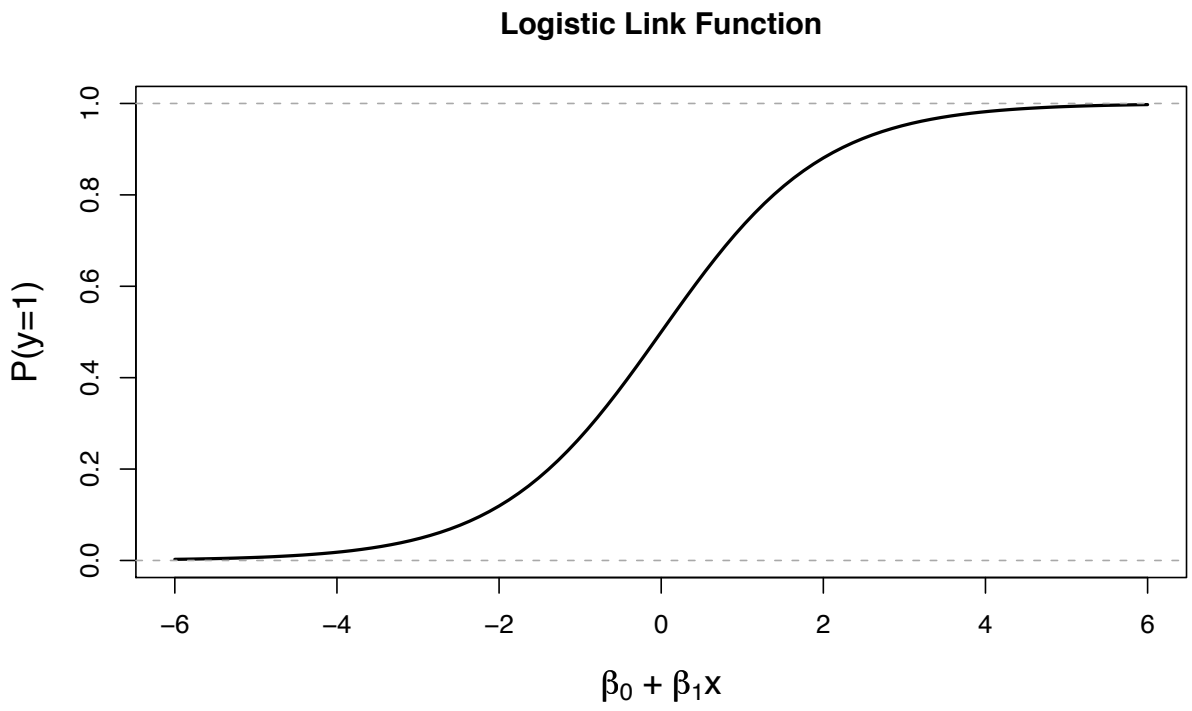

Figure 5.2: The logistic link function.

the probability that $y_i = 1$, given x_i, then

$$
\begin{aligned}
p_i &= \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \\
p_i + p_i e^{\beta_0+\beta_1 x_i} &= e^{\beta_0+\beta_1 x_i} \\
p_i &= (1-p_i)e^{\beta_0+\beta_1 x_i} \\
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i
\end{aligned}
$$

Since $p_i = P(y_i = 1 \mid x_i$, we know that $1 - p_i = P(y_i = 0 \mid x_i)$. Therefore, the ratio $p_i/(1-p_i)$ is the odds in favor of the proposition that $y_i = 1$, given the predictor x_i. This means that the $\beta_0 + \beta_1 x_i$ is a linear predictor for the logarithm of the odds in favor of success ($y_i = 1$). This is

Estimating the parameters of the logistic regression model

In previous chapters we learned how to estimate the parameters of a linear regression model using the least-squares criterion. This involved choosing values of the regression parameters to minimize the quantity

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,,$$

where $\hat{y}_i$ is the value for y_i predicted by the regression equation.

In logistic regression, the analogue of least-squares is Gauss's principle of maximum likelihood. The idea here is to choose values for β_0 and β_1 that make the observed patterns of 1's and 0's as small a miracle as possible.

To understand how this works, observe the following two facts:

- If $y_i = 1$, then we have observed an event that occurred with probability $P(y_i = 1 \mid x_i)$. Under the logistic-regression model, we can write this probability as

$$P(y_i = 1 \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

- If $y_i = 0$, then we have observed an event that occurred with probability $P(y_i = 0 \mid x_i) = 1 - P(y_i = 1 \mid x_i)$. Under the logistic regression model, we can write this probability as

$$1 - P(y_i = 1 \mid x_i) = 1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Since all of the individual 1's and 0's are independent, given the parameters β_0 and β_1, the probability of having observed our entire data set is the product of the probabilities for the individual 1's and 0's. We can write this as:

$$P(y_1, \ldots, y_n) = \prod_{i:y_i=1}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right) \cdot \prod_{i:y_i=0}\left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right).$$

This expression is our *likelihood*; it is the probability of having observed our data, given some particular configuration of the model parameters.[1] The logic of maximum likelihood is to choose values for β_0 and β_1 such that $P(y_1, \ldots, y_n)$ is as large as possible. We denote these choices by $\widehat{\beta}_0$ and $\widehat{\beta}_1$. These are called the *maximum-likelihood estimates* (MLE's) for the logistic regression model.

[1] The big $\prod$ signs mean "product," just like $\sum$ means "sum." The first product is for the observations where y_i was a 1, and the second product is for the observations where y_i was a 0.

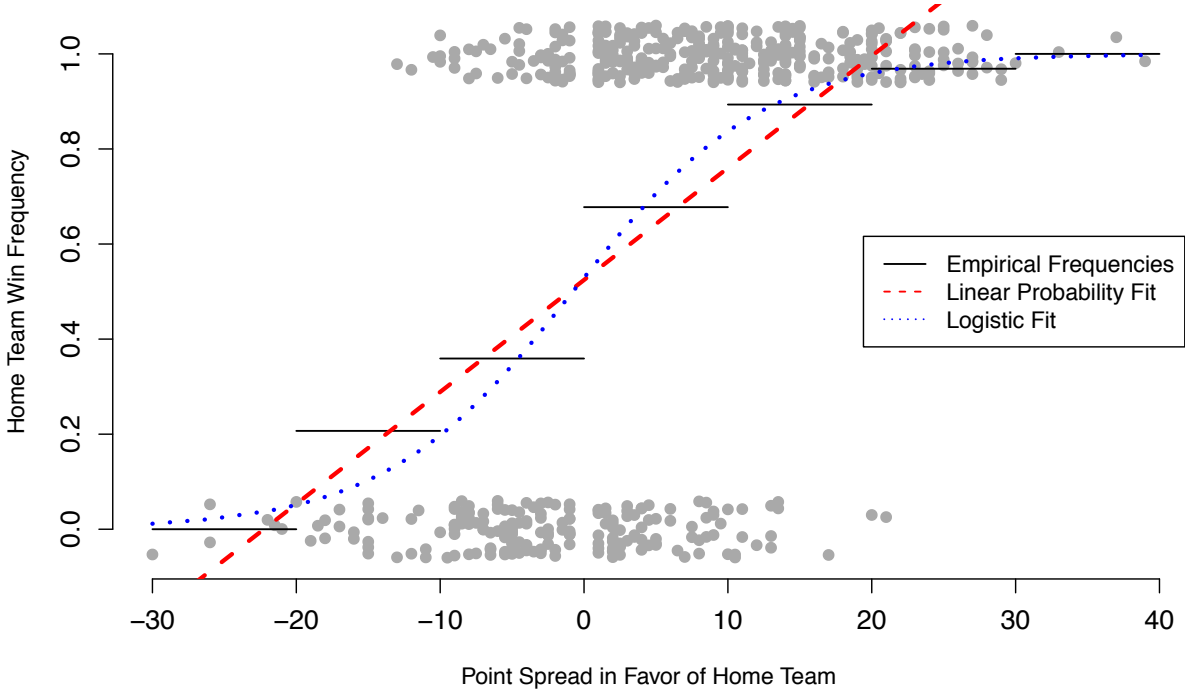

Figure 5.3: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

This likelihood is a difficult expression to maximize by hand (i.e. using calculus and algebra). Luckily, most major statistical software packages have built-in routines for fitting logistic-regression models, absolving you of the need to do any difficult analytical work.

The logistic regression fit for the point-spread data

Let's return briefly to the data on point spreads in NCAA basketball games. The figure above compares the logistic model to the linear-probability model. The logistic regression fit ($\widehat{\beta}_0 = 0.117$, $\widehat{\beta}_1 = 0.152$) eliminates the undesirable behavior of the linear model, and ensures that all forecasted probabilities are between 0 and 1. Note the clearly non-linear behavior of the dotted blue curve. Instead of fitting a straight line to the empirical success frequencies, we have fit an S-shape.