CHAPTER 5

Logistic regression

Logistic regression is the standard way to model binary outcomes (that is, data y_i that take on the values 0 or 1). Section 5.1 introduces logistic regression in a simple example with one predictor, then for most of the rest of the chapter we work through an extended example with multiple predictors and interactions.

5.1 Logistic regression with a single predictor

Example: modeling political preference given income

Conservative parties generally receive more support among voters with higher incomes. We illustrate classical logistic regression with a simple analysis of this pattern from the National Election Study in 1992. For each respondent i in this poll, we label $y_i = 1$ if he or she preferred George Bush (the Republican candidate for president) or 0 if he or she preferred Bill Clinton (the Democratic candidate), for now excluding respondents who preferred Ross Perot or other candidates, or had no opinion. We predict preferences given the respondent's income level, which is characterized on a five-point scale.[1]

The data are shown as (jittered) dots in Figure 5.1, along with the fitted *logistic regression* line, a curve that is constrained to lie between 0 and 1. We interpret the line as the probability that $y = 1$ given x—in mathematical notation, $\Pr(y = 1|x)$.

We fit and display the logistic regression using the following R function calls:

R code

```
fit.1 <- glm (vote ~ income, family=binomial(link="logit"))
display (fit.1)
```

to yield

R output

```
            coef.est coef.se
(Intercept)    -1.40    0.19
income          0.33    0.06
  n = 1179, k = 2
  residual deviance = 1556.9, null deviance = 1591.2 (difference = 34.3)
```

The fitted model is $\Pr(y_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$. We shall define this model mathematically and then return to discuss its interpretation.

The logistic regression model

It would not make sense to fit the continuous linear regression model, $X\beta + \text{error}$, to data y that take on the values 0 and 1. Instead, we model the probability that $y = 1$,

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta), \tag{5.1}$$

under the assumption that the outcomes y_i are independent given these probabilities. We refer to $X\beta$ as the *linear predictor*.

[1] See Section 4.7 for details on the income categories and other variables measured in this survey.

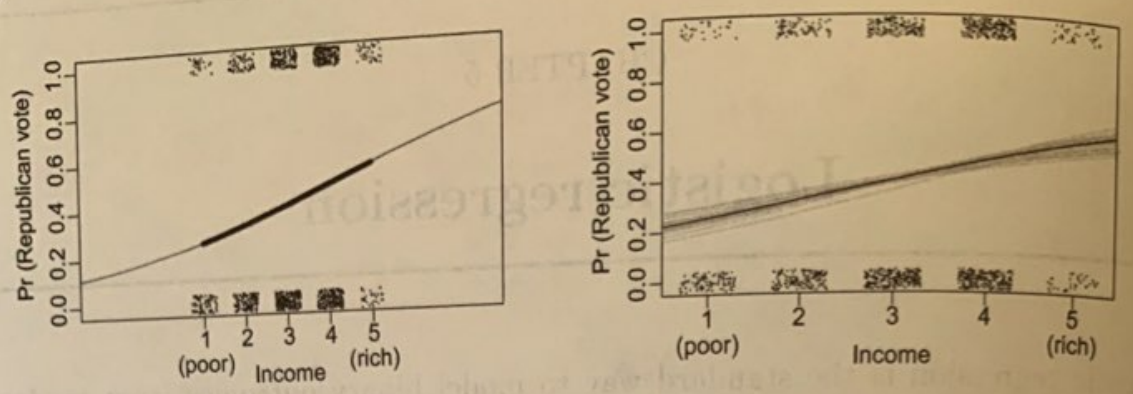

Figure 5.1 *Logistic regression estimating the probability of supporting George Bush in the 1992 presidential election, as a function of discretized income level. Survey data are indicated by jittered dots. In this example little is revealed by these jittered points, but we want to emphasize here that the data and fitted model can be put on a common scale. (a) Fitted logistic regression: the thick line indicates the curve in the range of the data; the thinner lines at the end show how the logistic curve approaches 0 and 1 in the limits. (b) In the range of the data, the solid line shows the best-fit logistic regression, and the light lines show uncertainty in the fit.*

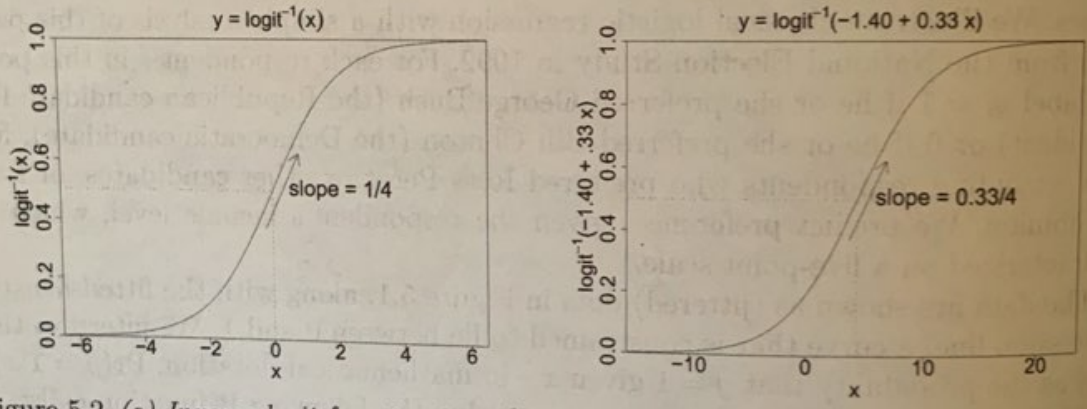

Figure 5.2 *(a) Inverse-logit function $\text{logit}^{-1}(x)$: the transformation from linear predictors to probabilities that is used in logistic regression. (b) An example of the predicted probabilities from a logistic regression model: $y = \text{logit}^{-1}(-1.40+0.33x)$. The shape of the curve is the same, but its location and scale have changed; compare the x-axes on the two graphs. For each curve, the dotted line shows where the predicted probability is 0.5: in graph (a), this is at $\text{logit}(0.5) = 0$; in graph (b), the halfway point is where $-1.40+0.33x = 0$, which is $x = 1.40/0.33 = 4.2$.*

The slope of the curve at the halfway point is the logistic regression coefficient divided by 4, thus 1/4 for $y = \text{logit}^{-1}(x)$ and 0.33/4 for $y = \text{logit}^{-1}(-1.40+0.33x)$. The slope of the logistic regression curve is steepest at this halfway point.

The function $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ transforms continuous values to the range $(0, 1)$, which is necessary, since probabilities must be between 0 and 1. This is illustrated for the election example in Figure 5.1 and more theoretically in Figure 5.2.

Equivalently, model (5.1) can be written

$$\begin{aligned} \Pr(y_i = 1) &= p_i \\ \text{logit}(p_i) &= X_i\beta, \end{aligned} \tag{5.2}$$

where $\text{logit}(x) = \log(x/(1-x))$ is a function mapping the range $(0, 1)$ to the range $(-\infty, \infty)$. We prefer to work with logit^{-1} because it is natural to focus on the mapping from the linear predictor to the probabilities, rather than the reverse. However, you will need to understand formulation (5.2) to follow the literature and also when fitting logistic models in Bugs.

The inverse-logistic function is curved, and so the expected difference in y corresponding to a fixed difference in x is not a constant. As can be seen in Figure 5.2, the steepest change occurs at the middle of the curve. For example:

- $\text{logit}(0.5) = 0$, and $\text{logit}(0.6) = 0.4$. Here, adding 0.4 on the logit scale corresponds to a change from 50% to 60% on the probability scale.
- $\text{logit}(0.9) = 2.2$, and $\text{logit}(0.93) = 2.6$. Here, adding 0.4 on the logit scale corresponds to a change from 90% to 93% on the probability scale.

Similarly, adding 0.4 at the low end of the scale moves a probability from 7% to 10%. In general, any particular change on the logit scale is compressed at the ends of the probability scale, which is needed to keep probabilities bounded between 0 and 1.

5.2 Interpreting the logistic regression coefficients

Coefficients in logistic regression can be challenging to interpret because of the nonlinearity just noted. We shall try to generalize the procedure for understanding coefficients one at a time, as was done for linear regression in Chapter 3. We illustrate with the model, $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$. Figure 5.1 shows the story, but we would also like numerical summaries. We present some simple approaches here and return in Section 5.7 to more comprehensive numerical summaries.

Evaluation at and near the mean of the data

The curve of the logistic function requires us to choose where to evaluate changes, if we want to interpret on the probability scale. The mean of the input variables in the data is often a useful starting point.

- As with linear regression, the *intercept* can only be interpreted assuming zero values for the other predictors. When zero is not interesting or not even in the model (as in the voting example, where income is on a 1–5 scale), the intercept must be evaluated at some other point. For example, we can evaluate Pr(Bush support) at the central income category and get $\text{logit}^{-1}(-1.40 + 0.33 \cdot 3) = 0.40$.

  Or we can evaluate Pr(Bush support) at the mean of respondents' incomes: $\text{logit}^{-1}(-1.40 + 0.33 \cdot \bar{x})$; in *R* we code this as[2]

  ```
  invlogit (-1.40 + 0.33*mean(income))
  ```
  R code

  or, more generally,

  ```
  invlogit (coef(fit.1)[1] + coef(fit.1)[2]*mean(income))
  ```
  R code

  For this dataset, $\bar{x} = 3.1$, yielding $\Pr(\text{Bush support}) = 0.40$ at this central point.
- A difference of 1 in income (on this 1–5 scale) corresponds to a positive difference of 0.33 in the logit probability of supporting Bush. There are two convenient ways to summarize this directly in terms of probabilities.
  - We can evaluate how the probability differs with a unit difference in x near the central value. Since $\bar{x} = 3.1$ in this example, we can evaluate the logistic regression function at $x = 3$ and $x = 2$; the difference in $\Pr(y = 1)$ corresponding to adding 1 to x is $\text{logit}^{-1}(-1.40+0.33\cdot 3) - \text{logit}^{-1}(-1.40+0.33\cdot 2) = 0.08$.

[2] We are using a function we have written, `invlogit <- function (x) {1/(1+exp(-x))}`.

A difference of 1 in income category corresponds to a positive difference of 8% in the probability of supporting Bush.

- Rather than consider a discrete change in x, we can compute the derivative of the logistic curve at the central value, in this case $\bar{x} = 3.1$. Differentiating the function $\text{logit}^{-1}(\alpha + \beta x)$ with respect to x yields $\beta e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})^2$. The value of the linear predictor at the central value of $\bar{x} = 3.1$ is $-1.40+0.33 \cdot 3.1 = -0.39$, and the slope of the curve—the "change" in $\Pr(y=1)$ per small unit of "change" in x—at this point is $0.33e^{-0.39}/(1 + e^{-0.39})^2 = 0.13$.
- For this example, the difference on the probability scale is the same value of 0.13 (to one decimal place); this is typical but in some cases where a unit difference is large, the differencing and the derivative can give slightly different answers. They will always be the same sign, however.

The "divide by 4 rule"

The logistic curve is steepest at its center, at which point $\alpha + \beta x = 0$ so that $\text{logit}^{-1}(\alpha+\beta x) = 0.5$ (see Figure 5.2). The slope of the curve—the derivative of the logistic function—is maximized at this point and attains the value $\beta e^0/(1+e^0)^2 = \beta/4$. Thus, $\beta/4$ is the maximum difference in $\Pr(y = 1)$ corresponding to a unit difference in x.

As a rule of convenience, we can take logistic regression coefficients (other than the constant term) and divide them by 4 to get an upper bound of the predictive difference corresponding to a unit difference in x. This upper bound is a reasonable approximation near the midpoint of the logistic curve, where probabilities are close to 0.5.

For example, in the model $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$, we can divide 0.33/4 to get 0.08: a difference of 1 in income category corresponds to no more than an 8% positive difference in the probability of supporting Bush. Because the data in this case actually lie near the 50% point (see Figure 5.1), this "divide by 4" approximation turns out to be close to 0.13, the derivative evaluated at the central point of the data.

Interpretation of coefficients as odds ratios

Another way to interpret logistic regression coefficients is in terms of *odds ratios*. If two outcomes have the probabilities $(p, 1-p)$, then $p/(1-p)$ is called the *odds*. An odds of 1 is equivalent to a probability of 0.5—that is, equally likely outcomes. Odds of 0.5 or 2.0 represent probabilities of $(1/3, 2/3)$. The ratio of two odds—thus, $(p_1/(1-p_1))/(p_2/(1-p_2))$—is called an odds ratio. Thus, an odds ratio of 2 corresponds to a change from $p = 0.33$ to $p = 0.5$, or a change from $p = 0.5$ to $p = 0.67$.

An advantage of working with odds ratios (instead of probabilities) is that it is possible to keep scaling up odds ratios indefinitely without running into the boundary points of 0 and 1. For example, going from an odds of 2 to an odds of 4 increases the probability from 2/3 to 4/5; doubling the odds again increases the probability to 8/9, and so forth.

Exponentiated logistic regression coefficients can be interpreted as odds ratios. For simplicity, we illustrate with a model with one predictor, so that

$$\log\left(\frac{\Pr(y=1|x)}{\Pr(y=0|x)}\right) = \alpha + \beta x. \tag{5.3}$$

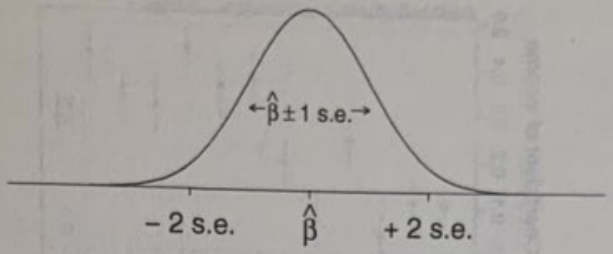

Figure 5.3 *Distribution representing uncertainty in an estimated regression coefficient (repeated from page 40). The range of this distribution corresponds to the possible values of β that are consistent with the data. When using this as an uncertainty distribution, we assign an approximate 68% chance that β will lie within 1 standard error of the point estimate, $\hat{\beta}$, and an approximate 95% chance that β will lie within 2 standard errors. Assuming the regression model is correct, it should happen only about 5% of the time that the estimate, $\hat{\beta}$, falls more than 2 standard errors away from the true β.*

Adding 1 to x (that is, changing x to $x+1$ in (5.3)) has the effect of adding β to both sides of the equation. Exponentiating both sides, the odds are then multiplied by e^{β}. For example, if $\beta = 0.2$, then a unit difference in x corresponds to a multiplicative change of $e^{0.2} = 1.22$ in the odds (for example, changing the odds from 1 to 1.22, or changing p from 0.5 to 0.55).

We find that the concept of odds can be somewhat difficult to understand, and odds ratios are even more obscure. Therefore we prefer to interpret coefficients on the original scale of the data when possible, for example, saying that adding 0.2 on the logit scale corresponds to a change in probability from $\text{logit}^{-1}(0)$ to $\text{logit}^{-1}(0.2)$.

Inference

Coefficient estimates and standard errors. The coefficients in classical logistic regression are estimated using maximum likelihood, a procedure that can often work well for models with few predictors fit to reasonably large samples (but see Section 5.8 for a potential problem).

As with the linear model, the standard errors represent estimation uncertainty. We can roughly say that coefficient estimates within 2 standard errors of $\hat{\beta}$ are consistent with the data. Figure 5.3 shows the normal distribution that approximately represents the range of possible values of β. For the voting example, the coefficient of income has an estimate $\hat{\beta}$ of 0.33 and a standard error of 0.06; thus the data are roughly consistent with values of β in the range $[0.33 \pm 2 \cdot 0.06] = [0.21, 0.45]$.

Statistical significance. As with linear regression, a coefficient is considered "statistically significant" if it is at least 2 standard errors away from zero. In the voting example, the coefficient of income is statistically significant and positive, meaning that we can be fairly certain that, in the population represented by this survey, positive differences in income generally correspond to positive (not negative) differences in the probability of supporting Bush for president.

Also as with linear regression, we usually do *not* try to interpret the statistical significance of the intercept. The sign of an intercept is not generally of any interest, and so it is usually meaningless to compare it to zero or worry about whether it is statistically significantly different from zero.

Finally, when considering multiple inputs, we follow the same principles as with linear regression when deciding when and how to include and combine inputs in a model, as discussed in Section 4.6.