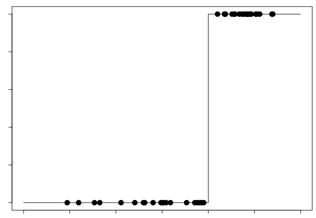

Figure 5.16 *Example of data for which a logistic regression model is nonidentifiable. The outcome* y *equals* 0 *for all data below* $x = 2$ *and* 1 *for all data above* $x = 2$*, hence the best-fit logistic regression line is* $y = \text{logit}^{-1}(\infty(x-2))$*, which has an infinite slope at* $x = 2$.

education) and predictors (constant term, distance, arsenic, education, and distance $\times$ arsenic) is crucial. We discuss average predictive comparisons further in Section 21.4.

5.8 Identifiability and separation

There are two reasons that a logistic regression can be nonidentified (that is, have parameters that cannot be estimated from the available data and model, as discussed in Section 4.5 in the context of linear regression):

1. As with linear regression, if predictors are collinear, then estimation of the linear predictor, $X\beta$, does not allow separate estimation of the individual parameters β. We can handle this kind of nonidentifiability in the same way that we would proceed for linear regression, as described in Section 4.5.

2. A completely separate identifiability problem, called *separation*, can arise from the discreteness of the data.

   - If a predictor x_j is completely aligned with the outcome, so that $y = 1$ for all the cases where x_j exceeds some threshold T, and $y = 0$ for all cases where $x_j < T$, then the best estimate for the coefficient β_j is ∞. Figure 5.16 shows an example. Exercise 5.11 gives an example with a binary predictor.

   - Conversely, if $y = 1$ for all cases where $x_j < T$, and $y = 0$ for all cases where $x_j > T$, then $\hat{\beta}_j$ will be $-\infty$.

   - More generally, this problem will occur if any linear combination of predictors is perfectly aligned with the outcome. For example, suppose that $7x_1 + x_2 - 3x_3$ is completely positively aligned with the data, with $y = 1$ if and only if this linear combination of predictors exceeds some threshold. Then the linear combination $7\hat{\beta}_1 + \hat{\beta}_2 - 3\hat{\beta}_3$ will be estimated at ∞, which will cause at least one of the three coefficients $\beta_1, \beta_2, \beta_3$ to be estimated at ∞ or $-\infty$.

   One way to handle separation is using a Bayesian or penalized-likelihood approach (implemented for R in the `brlr` package) that provides a small amount of information on all the regression coefficients, including those that are not identified from the data alone. (See Chapter 18 for more on Bayesian inference.)