Example: Hypothetical rare disease

We illustrate with a simple example in which models are fit to predict the occurrence or non-occurrence of a rare disease that has 1% prevalence in the dataset and also in the population. Each model will then be used to make predictions out of sample. Consider the following two models:

1. Logistic regression with only a constant term. The predicted probability of disease is then 0.01 for each person.
2. Logistic regression with several useful predictors (for example, age, sex, and some measures of existing symptoms). Suppose that, under this model, predictive probabilities are very low (for example, in the range of 0.001 or 0.0001) for most people, but for some small subgroup, the model will identify them as high risk and give them a 0.40 predictive probability of having the disease.

Model 2, assuming it fits the data well and that the sample is representative of the population, is much better than model 1, as it successfully classifies some people as being of high disease risk, while reducing worry about the vast majority of the population.

But now consider comparing the two models using error rate:

1. The simple model, which assigns probability 0.01 to everyone, will give "no disease" as its point prediction for every person. We have stipulated that 1% of people actually have the disease, so the error rate is 1%.
2. The more complex model still happens to return predictive probability of disease at less then 50% for all people, so it still gives a point prediction of "no disease" for every person, and the error rate is still 1%.

Model 2, despite being much more accurate, and identifying some people to have as high as a 40% chance of having the disease, still gives no improvement in error rate. This does not mean that Model 2 is no better than Model 1; it just tells us that error rate is an incomplete summary.

14.6 Identification 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 10.7 in the context of linear regression):

- 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 10.7.
- 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 14.11 shows an example.
  - 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 approach that provides some amount of information on all the regression coefficients, including those that are not identified from the data alone.

When a fitted model "separates" (that is, perfectly predicts some subset of) discrete data, the maximum likelihood estimate can be undefined or unreasonable, and we can get a more reasonable

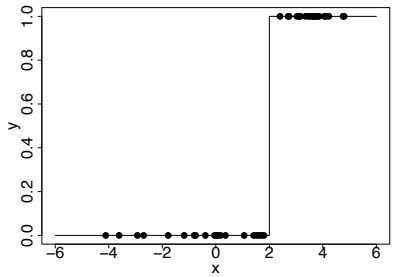

Figure 14.11 *Example of data for which a logistic regression model is nonidentifiable. The outcome y equals 0 for all data below $x = 2$ and 1 for all data above $x = 2$; hence the best-fit logistic regression line is $y = \text{logit}^{-1}(\infty(x-2))$, which has an infinite slope at $x = 2$.*

```
1960
            coef.est coef.se
(Intercept)  -0.14     0.23
female        0.24     0.14
black        -1.03     0.36
income        0.03     0.06
  n = 875, k = 4

1964
            coef.est coef.se
(Intercept)  -1.15     0.22
female       -0.08     0.14
black       -16.83   420.40
income        0.19     0.06
  n = 1058, k = 4

1968
            coef.est coef.se
(Intercept)   0.26     0.24
female       -0.03     0.14
black        -3.53     0.59
income        0.00     0.07
  n = 876, k = 4

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
  n = 1518, k = 4
```

Figure 14.12 *Estimates and standard errors from logistic regressions (maximum likelihood estimates, which roughly correspond to Bayesian estimates with uniform prior distributions) predicting Republican vote intention in pre-election polls, fit separately to survey data from four presidential elections from 1960 through 1972. The estimates are reasonable except in 1964, where there is complete separation (with none of black respondents supporting the Republican candidate, Barry Goldwater), which leads to an essentially infinite estimate of the coefficient for* `black`.

answer using prior information to effectively constrain the inference. We illustrate with an example that arose in one of our routine analyses, where we were fitting logistic regressions to a series of datasets.

Example: Income and voting

Figure 14.12 shows the estimated coefficients in a model predicting probability of Republican vote for president given sex, ethnicity, and income (coded as $-2, -1, 0, 1, 2$), fit separately to pre-election polls for a series of elections.[3] The estimates look fine except in 1964, where there is complete separation: of the 87 African Americans in the survey that year, none reported a preference for the Republican candidate. We fit the model in R, which actually yielded a finite estimate for the coefficient of `black` even in 1964, but that number and its standard error are essentially meaningless, being a function of how long the iterative fitting procedure goes before giving up. The maximum likelihood estimate for the coefficient of `black` in that year is $-\infty$.

Figure 14.13 displays the *profile likelihood* for the coefficient of `black` in the voting example, that is, the maximum value of the likelihood function conditional on this coefficient. The maximum of the profile likelihood is at $-\infty$, a value that makes no sense in this application. As we shall see, an informative prior distribution will bring the estimate down to a reasonable value.

[3]Data and code for this example are in the folder `NES`.

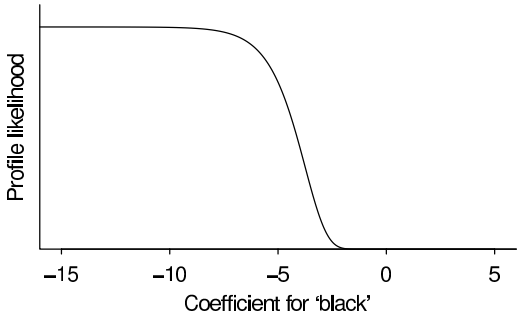

Figure 14.13 *Profile likelihood (in this case, essentially the posterior density given a uniform prior distribution) of the coefficient of* `black` *from the logistic regression of Republican vote in 1964 (displayed in the lower left of Figure 14.12), conditional on point estimates of the other coefficients in the model. The maximum occurs as* $\beta \to -\infty$*, indicating that the best fit to the data would occur at this unreasonable limit. The* y*-axis starts at 0; we do not otherwise label it because all that matters are the relative values of this curve, not its absolute level.*

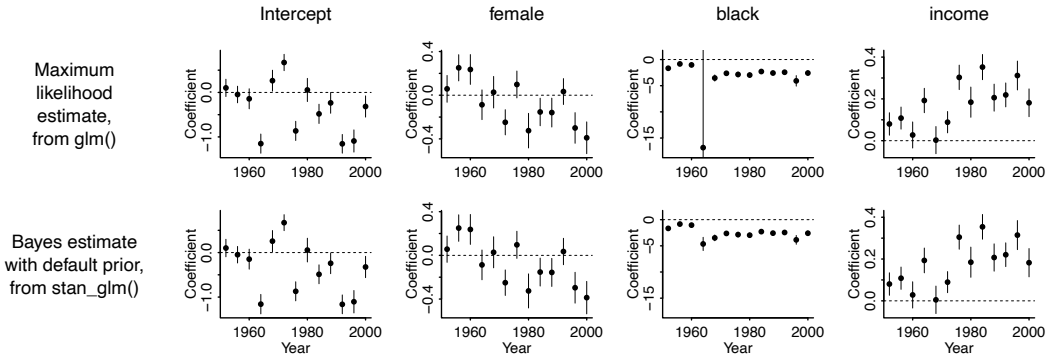

Figure 14.14 *The top row shows the estimated coefficients (*± 1 *standard error) for a logistic regression predicting probability of Republican vote for president, given sex, race, and income, as fit separately to data from the National Election Study for each election 1952 through 2000. The complete separation in 1964 led to a coefficient estimate of* $-\infty$ *that year. (The particular finite values of the estimate and standard error are determined by the number of iterations used by the* `glm` *function in R before stopping.) The bottom row shows the posterior estimates and standard deviations for the same model fit each year using Bayesian inference with default priors from* `stan_glm`*. The Bayesian procedure does a reasonable job at stabilizing the estimates for 1964, while leaving the estimates for other years essentially unchanged.*

We apply our default prior distribution to the pre-election polls discussed earlier in this section, in which we could easily fit a logistic regression with flat priors for every election year except 1964, where the coefficient for `black` blew up due to complete separation in the data.

The top row of Figure 14.14 shows the time series of estimated coefficients and error bounds for the four coefficients of the logistic regression fit using `glm` separately to poll data for each election year from 1952 through 2000. All the estimates are reasonable except for the coefficient for `black` in 1964, when the maximum likelihood estimates are infinite. As discussed already, we do not believe that these coefficient estimates for 1964 are reasonable: in the population as a whole, we do *not* believe that the probability was zero that an African American in the population would vote Republican. It is, however, completely predictable that with moderate sample sizes there will occasionally be separation, yielding infinite estimates. (The estimates from `glm` shown here for 1964 are finite only because the generalized linear model fitting routine in R stops after a finite number of iterations.)

The bottom row of Figure 14.14 shows the coefficient estimates fit using `stan_glm` on its default setting. The estimated coefficient for `black` in 1964 has been stabilized, with the other coefficients

being essentially unchanged. This example illustrates how a weakly informative prior distribution can work in routine practice.

Weakly informative default prior compared to actual prior information

The default prior distribution used by `stan_glm` does not represent our prior knowledge about the coefficient for `black` in the logistic regression for 1964 or any other year. Even before seeing data from this particular series of surveys, we knew that blacks have been less likely than whites to vote for Republican candidates; thus our prior belief was that the coefficient was negative. Furthermore, our prior for any given year would be informed by data from other years. For example, given the series of estimates in Figure 14.14, we would guess that the coefficient for `black` in 2004 is probably between -2 and -5. Finally, we are not using specific prior knowledge about these elections. The Republican candidate in 1964 was particularly conservative on racial issues and opposed the Civil Rights Act; on those grounds we would expect him to poll particularly poorly among African Americans (as indeed he did). In sum, we feel comfortable using a default model that excludes much potentially useful information, recognizing that we could add such information if it were judged to be worth the trouble (for example, instead of performing separate estimates for each election year, setting up a hierarchical model allowing the coefficients to gradually vary over time and including election-level predictors including information such as the candidates' positions on economic and racial issues).

14.7 Bibliographic note

Binned residual plots and related tools for checking the fit of logistic regressions are discussed by Landwehr, Pregibon, and Shoemaker (1984); Gelman, Goegebeur, et al. (2000); and Pardoe and Cook (2002).

See Gelman and Pardoe (2007) for more on average predictive comparisons and Gelman, Goegebeur et al. (2000) for more on binned residuals.

Nonidentifiability of logistic regression and separation in discrete data are discussed by Albert and Anderson (1984), Lesaffre and Albert (1989), Heinze and Schemper (2003), as well as in the book by Agresti (2002). Firth (1993), Zorn (2005), and Gelman et al. (2008) present Bayesian solutions.

14.8 Exercises

14.1 *Graphing binary data and logistic regression*: Reproduce Figure 14.1 with the model, $\Pr(y = 1) = \text{logit}^{-1}(0.4 - 0.3x)$, with 50 data points x sampled uniformly in the range $[A, B]$. (In Figure 14.1 the x's were drawn from a normal distribution.) Choose the values A and B so that the plot includes a zone where values of y are all 1, a zone where they are all 0, and a band of overlap in the middle.

14.2 *Logistic regression and discrimination lines*: Reproduce Figure 14.2 with the model, $\Pr(y = 1) = \text{logit}^{-1}(0.4 - 0.3x_1 + 0.2x_2)$, with (x_1, x_2) sampled uniformly from the rectangle $[A_1, B_1] \times [A_2, B_2]$. Choose the values A_1, B_1, A_2, B_2 so that the plot includes a zone where values of y are all 1, a zone where they are all 0, and a band of overlap in the middle, and with the three lines corresponding to $\Pr(y = 1) = 0.1$, 0.5, and 0.9 are all visible.

14.3 *Graphing logistic regressions*: The well-switching data described in Section 13.7 are in the folder `Arsenic`.

(a) Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

(b) Make a graph similar to Figure 13.8b displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

(c) Make a residual plot and binned residual plot as in Figure 14.8.