CHAPTER 14

Multilevel logistic regression

Multilevel modeling is applied to logistic regression and other generalized linear models in the same way as with linear regression: the coefficients are grouped into batches and a probability distribution is assigned to each batch. Or, equivalently (as discussed in Section 12.5), error terms are added to the model corresponding to different sources of variation in the data. We shall discuss logistic regression in this chapter and other generalized linear models in the next.

14.1 State-level opinions from national polls

Dozens of national opinion polls are conducted by media organizations before every election, and it is desirable to estimate opinions at the levels of individual states as well as for the entire country. These polls are generally based on national random-digit dialing with corrections for nonresponse based on demographic factors such as sex, ethnicity, age, and education.

Here we describe a model developed for estimating state-level opinions from national polls, while simultaneously correcting for nonresponse, for any survey response of interest. The procedure has two steps: first fitting the model and then applying the model to estimate opinions by state:

1. We fit a regression model for the individual response y given demographics and state. This model thus estimates an average response θ_l for each cross-classification l of demographics and state. In our example, we have sex (male or female), ethnicity (African American or other), age (4 categories), education (4 categories), and 51 states (including the District of Columbia); thus $l = 1, \ldots, L = 3264$ categories.
2. From the U.S. Census, we look up the adult population N_l for each category l. The estimated population average of the response y in any state j is then

$$\theta_j = \sum_{l \in j} N_l \theta_l \Big/ \sum_{l \in j} N_l, \tag{14.1}$$

with each summation over the 64 demographic categories l in the state. This weighting by population totals is called *poststratification* (see the footnote on page 181). In the actual analysis we also considered poststratification over the population of eligible voters but we do not discuss this further complication here.

We need many categories because (a) we are interested in estimates for individual states, and (b) nonresponse adjustments force us to include the demographics. As a result, any given survey will have few or no data in many categories. This is not a problem, however, if a multilevel model is fitted. Each factor or set of interactions in the model is automatically given a variance component. This inferential procedure works well and outperforms standard survey estimates when estimating state-level outcomes.

In this demonstration, we choose a single outcome—the probability that a respondent prefers the Republican candidate for president—as estimated by a logistic

regression model from a set of seven CBS News polls conducted during the week before the 1988 presidential election.

A simple model with some demographic and geographic variation

We label the survey responses y_i as 1 for supporters of the Republican candidate and 0 for supporters of the Democrat (with undecideds excluded) and model them as independent, with $\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$. Potential input variables include the state index $j[i]$ and the demographics used by CBS in the survey weighting: categorical variables for sex, ethnicity, age, and education.

We introduce multilevel logistic regression with a simple example including two individual predictors—`female` and `black`—and the 51 states:

$$\begin{aligned} \Pr(y_i=1) &= \text{logit}^{-1}\left(\alpha_{j[i]} + \beta^{\text{female}} \cdot \text{female}_i + \beta^{\text{black}} \cdot \text{black}_i\right), \text{ for } i = 1, \ldots, n \\ \alpha_j &\sim \text{N}\left(\mu_\alpha, \sigma^2_{\text{state}}\right), \text{ for } j = 1, \ldots, 51. \end{aligned}$$

We can quickly fit the model in R,

R code
```
M1 <- lmer(y ~ black + female + (1|state), family=binomial(link="logit"))
display (M1)
```

and get the following:

R code
```
             coef.est coef.se
(Intercept)  0.4      0.1
black       -1.7      0.2
female      -0.1      0.1
Error terms:
 Groups          Name        Std.Dev.
 state           (Intercept) 0.4
 No residual sd
# of obs: 2015, groups: state, 49
deviance = 2658.7
  overdispersion parameter = 1.0
```

The top part of this display gives the estimate of the average intercept, the coefficients for `black` and `female`, and their standard errors. Reading down, we see that σ_{state} is estimated at 0.4. There is no "residual standard deviation" because the logistic regression model does not have such a parameter (or, equivalently, it is fixed to the value 1.6, as discussed near the end of Section 5.3). The deviance (see page 100) is printed as a convenience but we usually do not look at it. Finally, the model has an overdispersion of 1.0—that is, no overdispersion—because logistic regression with binary data (as compared to count data; see Section 6.3) cannot be overdispersed.

We can also type `coef(M1)` to examine the estimates and standard errors of the state intercepts α_j, but rather than doing this we shall move to a larger model including additional predictors at the individual and state level. Recall that our ultimate goal here is not to estimate the α's, β's, and σ's, but to estimate the average value of y within each of the poststratification categories, and then to average over the population using the census numbers using equation (14.1).

A fuller model including non-nested factors

We expand the model to use all the demographic predictors used in the CBS weighting, including sex × ethnicity and age × education. We model age and education

(with four categories each) with varying intercepts, and also model the 16 levels of the age × education interaction.

At the state level, we include indicators for the 5 regions of the country (Northeast, Midwest, South, West, and D.C., considered as a separate region because of its distinctive voting patterns), along with `v.prev`, a measure of previous Republican vote in the state (more precisely, the average Republican vote share in the three previous elections, adjusted for home-state and home-region effects in the previous elections).

We shall write the model using indexes j, k, l, m for state, age category, education category, and region:

$$\begin{aligned}\Pr(y_i=1) &= \text{logit}^{-1}\big(\beta^0 + \beta^{\text{female}}\cdot\text{female}_i + \beta^{\text{black}}\cdot\text{black}_i + \\ &\quad + \beta^{\text{female.black}}\cdot\text{female}_i\cdot\text{black}_i + \alpha^{\text{age}}_{k[i]} + \alpha^{\text{edu}}_{l[i]} + \alpha^{\text{age.edu}}_{k[i],l[i]} + \alpha^{\text{state}}_{j[i]}\big) \\ \alpha^{\text{state}}_j &\sim \text{N}\left(\alpha^{\text{region}}_{m[j]} + \beta^{\text{v.prev}}\cdot\text{v.prev}_j,\ \sigma^2_{\text{state}}\right). \qquad (14.2)\end{aligned}$$

We also model the remaining multilevel coefficients:

$$\begin{aligned}\alpha^{\text{age}}_k &\sim \text{N}(0, \sigma^2_{\text{age}}),\ \text{for } k=1,\ldots,4 \qquad (14.3)\\ \alpha^{\text{edu}}_l &\sim \text{N}(0, \sigma^2_{\text{edu}}),\ \text{for } l=1,\ldots,4 \\ \alpha^{\text{age.edu}}_{k,l} &\sim \text{N}(0, \sigma^2_{\text{age.edu}}),\ \text{for } k=1,\ldots,4,\ l=1,\ldots,4 \\ \alpha^{\text{region}}_m &\sim \text{N}(0, \sigma^2_{\text{region}}),\ \text{for } m=1,\ldots,5. \qquad (14.4)\end{aligned}$$

As with the non-nested linear models in Section 13.5, this model can be expressed in equivalent ways by moving the constant term β_0 around. Here we have included β^0 in the data-level regression and included no intercepts in the group-level models for the different batches of α's.

Another approach is to include constant terms in several places in the model, centering the distributions in (14.4) at $\mu_{\text{age}}, \mu_{\text{edu}}, \mu_{\text{age}}, \mu_{\text{age.edu}}$, and μ_{region}. This makes the model nonidentifiable, but it can then be reparameterized in terms of identifiable combinations of parameters. Such a *redundant parameterization* speeds computation and offers some conceptual advantages, and we shall return to it in Section 19.4.

We can quickly fit model (14.2) in R: we first construct the index variable for the age × education interaction and expand the state-level predictors to the data level:

R code

```
age.edu <- n.edu*(age-1) + edu
region.full <- region[state]
v.prev.full <- v.prev[state]
```

We then fit and display the full multilevel model, to get:

R output

```
lmer(formula = y ~ black + female + black:female + v.prev.full +
    (1 | age) + (1 | edu) + (1 | age.edu) + (1 | state) +
    (1 | region.full), family = binomial(link = "logit"))
             coef.est coef.se
(Intercept)  -3.5      1.0
black        -1.6      0.3
female       -0.1      0.1
v.prev.full   7.0      1.7
black:female -0.2      0.4
Error terms:
 Groups         Name        Std.Dev.
```

```
 state          (Intercept) 0.2
 age.edu        (Intercept) 0.2
 region.full    (Intercept) 0.2
 edu            (Intercept) 0.1
 age            (Intercept) 0.0
 No residual sd
# of obs: 2015, groups: state,49; age.edu,16; region.full,5; edu,4; age,4
deviance = 2629.5
  overdispersion parameter = 1.0
```

Quickly reading this regression output:

- The intercept is not easily interpretable since it corresponds to a case in which `black`, `female`, and `v.prev` are all 0—but `v.prev` typically takes on values near 0.5 and is never 0.
- The coefficient for `black` is −1.6. Dividing by 4 (see page 82) yields a rough estimate that African-American men were 40% less likely than other men to support Bush, after controlling for age, education, and state.
- The coefficient for `female` is −0.1. Dividing by 4 yields a rough estimate that non-African-American women were very slightly less likely than non-African-American men to support Bush, after controlling for age, education, and state. However, the standard error on this coefficient is as large as the estimate itself, indicating that our sample size is too small for us to be certain of this pattern in the population.
- The coefficient for `v.prev.full` is 7.0, which, when divided by 4, is 1.7, suggesting that a 1% difference in a state's support for Republican candidates in previous elections mapped to a predicted 1.7% difference in support for Bush in 1988.
- The large standard error on the coefficient for `black:female` indicates that the sample size is too small to estimate this interaction precisely.
- The state-level errors have estimated standard deviation 0.2 on the logit scale. Dividing by 4 tells us that the states differed by approximately ±5% on the probability scale (over and above the differences explained by demographic factors).
- The differences among age-education groups and regions are also approximately ±5% on the probability scale.
- Very little variation is found among age groups or education groups after controlling for the other predictors in the model.

To make more precise inferences and predictions, we shall fit the model using Bugs (as described in Section 17.4), because with so many factors—including some with only 4 or 5 levels—the approximate inference provided by `lmer()` (which does not fully account for uncertainty in the estimated variance parameters) is not so reliable. It is still useful as a starting point, however, and we recommend performing the quick fit if possible before getting to more elaborate inference. In some other settings, it will be difficult to get Bugs to run successfully and we simply use the inferences from `lmer()`.

Graphing the estimated model

We would like to construct summary plots as we did with the multilevel models of Chapters 12 and 13. We alter the plotting strategy in two ways. First, the outcome is binary and so we plot $\Pr(y=1) = \mathrm{E}(y)$ as a function of the predictors; thus the graphs are curved, as are the classical generalized linear models in Chapter 6.

Our second modification of the plots is needed to deal with the many different predictors in our model: instead of plotting $\mathrm{E}(y)$ as a function of each of the demographic inputs, we combine them into a linear predictor for demographics, which we shall call $\mathrm{linpred}_i$:

$$\begin{aligned}\mathrm{linpred}_i &= \beta^0 + \beta^{\mathrm{female}} \cdot \mathrm{female}_i + \beta^{\mathrm{black}} \cdot \mathrm{black}_i + \\ &\quad + \beta^{\mathrm{female.black}} \cdot \mathrm{female}_i \cdot \mathrm{black}_i + \alpha^{\mathrm{age}}_{k[i]} + \alpha^{\mathrm{edu}}_{l[i]} + \alpha^{\mathrm{age.edu}}_{k[i],l[i]}. \quad (14.5)\end{aligned}$$

The estimates, 50% intervals, and 95% intervals for the demographic coefficients are displayed in Figure 14.1. Because all categories of each predictor variable have been included, these estimates can be interpreted directly as the contribution each makes to the sum, $X_i\beta$. So, for instance, if we were to predict the response for someone who is female, age 20, and with no high school diploma, we could simply take the constant term, plus the estimates for the corresponding three main effects plus the interaction between "18–29" and "no high school," plus the corresponding state coefficient, and then take the inverse-logit to obtain the probability of a Republican vote. As can be seen from the graph, the demographic factors other than ethnicity are estimated to have little predictive power. (Recall from Section 5.1 that we can quickly interpret logistic regression coefficients on the probability scale by dividing them by 4.)

For any survey respondent i, the regression prediction can then be written as

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\mathrm{linpred}_i + \alpha^{\mathrm{state}}_{j[i]}),$$

where $\mathrm{linpred}_i$ is the combined demographic predictor (14.5), and we can plot this for each state. We can do this in R—after first fitting the model in Bugs (as called from R) and attaching the resulting object, which puts arrays into the R workspace representing simulations for all the parameters from the model fit.

We summarize the linear predictor $\mathrm{linpred}_i$ from (14.5) by its average over the simulations. Recall that we are using simulations from the fitted model (see Section 17.4), which we shall call `M3.bugs`. As discussed in Chapter 16, the first step after fitting the model is to attach the Bugs object so that the vectors and arrays of parameter simulations can be accessed within the R workspace. Here is the code to compute the vector `linpred`:

R code

```
attach.bugs (M3.bugs)
linpred <- rep (NA, n)
for (i in 1:n){
  linpred[i] <- mean (b.0 + b.female*female[i] + b.black*black[i] +
    b.female.black*female[i]*black[i] + a.age[age[i]] + a.edu[edu[i]] +
    a.age.edu[age[i],edu[i]])
}
```

We can then make Figure 14.2 given the simulations from the fitted Bugs model:

R code

```
par (mfrow=c(2,4))
for (j in displayed.states){
  plot (0, 0, xlim=range(linpred), ylim=c(0,1), yaxs="i",
        xlab="linear predictor", ylab="Pr (support Bush)",
        main=state.name[j], type="n")
  for (s in 1:20){
    curve (invlogit (a.state[s,j] + x), lwd=.5, add=TRUE, col="gray")}
  curve (invlogit (median (a.state[,j]) + x), lwd=2, add=TRUE)
  if (sum(state==j)>0) points (linpred[state==j], y.jitter[state==j])
}
```

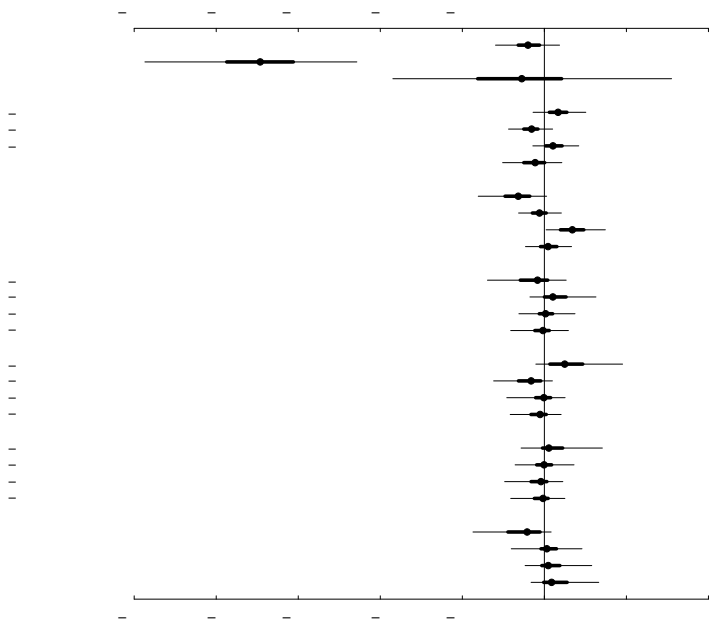

Figure 14.1 *Estimates, 50% intervals, and 95% intervals for the logistic regression coefficients for the demographic predictors in the model predicting the probability of supporting George Bush in polls before the 1988 presidential election. Recall that a change of x on the logistic scale corresponds to a change of at most $x/4$ on the probability scale. Thus, demographic factors other than ethnicity have small estimated predictive effects on vote preference.*

Figure 14.2 shows the result for a selection of eight states and illustrates a number of points about multilevel models. The solid lines display the estimated logistic regressions: thus, in any state, the probability of supporting Bush ranges from about 10% to 70% depending on the demographic variables—most importantly, ethnicity. Roughly speaking, there is about a 10% probability of supporting Bush for African Americans and about 60% for others, with other demographic variables slightly affecting the predicted probability. The variation among states is fairly small—you have to look at the different plots carefully to see it—but is important in allowing us to estimate average opinion by state, as we shall discuss. Changes of only a few percent in preferences can have large political impact.

The gray lines on the graphs represent uncertainty in the state-level coefficients, α_j^{state}. Alaska has no data at all, but the inference there is still reasonably precise—its α_j^{state} is estimated from its previous election outcome, its regional predictor (Alaska is categorized as a Western state), and from the distribution of the errors from the state-level regression. In general, the larger states such as California have more precise estimates than the smaller states such as Delaware—with more data in a state j, it is possible to estimate α_j^{state} more accurately.

The logistic regression curve is estimated for all states, even those such as Arizona with little range of x in the data (the survey included no black respondents from

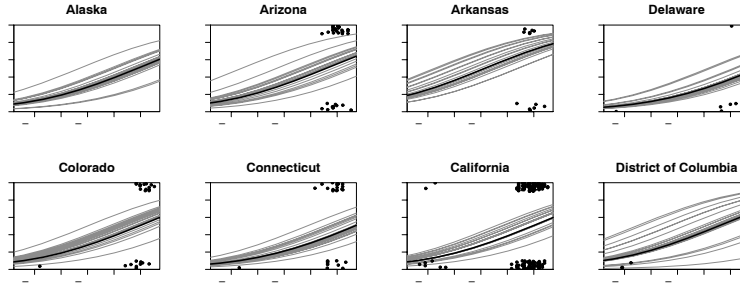

Figure 14.2 *Estimated probability of a survey respondent supporting Bush for president, as a function of the linear predictor for demographics, in each state (displaying only a selection of eight states, ordered by decreasing support for Bush, to save space). Dots show the data (y-jittered for visibility), and the heavy and light lines show the median estimate and 20 random simulation draws from the estimated model.*

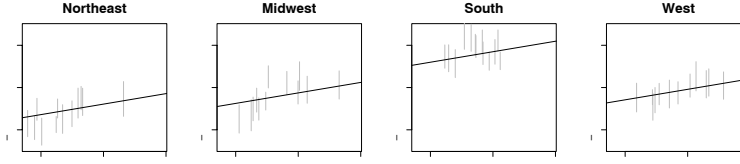

Figure 14.3 *Estimates and 50% intervals for the state coefficients* α_j^{state}*, plotted versus previous state vote* v.prev_j*, in each of the four regions of the United States. The estimated group-level regression line,* $\alpha_j^{\text{state}} = \alpha_{m[j]}^{\text{region}} + \beta_j^{\text{v.prev}} \cdot \text{v.prev}_j$*, is overlain on each plot (corresponding to regions* $m = 1, 2, 3, 4$*).*

Arizona). The model is set up so the demographic coefficients are the same for all states, so the estimate of the logistic curve is pooled for all the data. If the model included an interaction between demographics and state, then we would see differing slopes, and more uncertainty about the slope in states such as Arizona that have less variation in their data.

Figure 14.3 displays the estimated logistic regression coefficients for the 50 states, grouping them by region and, within each region, showing the state-level regression on `v.prev`, the measure of Republican vote in the state in previous presidential elections. Region and previous vote give good but not perfect predictions of the state-level coefficients in the public opinion model.

Using the model inferences to estimate average opinion for each state

The logistic regression model gives the probability that any adult will prefer Bush, given the person's sex, ethnicity, age, education level, and state. We can now compute weighted averages of these probabilities to represent the proportion of Bush supporters in any specified subset of the population.

We first extract from the U.S. Census the counts N_l in each of the 3264 cross-classification cells and create a 3264×6 data frame, `census`, indicating the sex,

```
     female black age edu state     N
1         0     0   1   1     1 66177
2         0     1   1   1     1 32465
3         1     0   1   1     1 59778
4         1     1   1   1     1 27416
5         0     0   2   1     1 83032
. . .
3262      0     1   4   4    51     5
3263      1     0   4   4    51  2610
3264      1     1   4   4    51     5
```

Figure 14.4 *The data frame* `census` *in R used for poststratification in the election polling example. The categories are ordered by ethnicity, sex, age category, education category, and state. The states are in alphabetical order; thus there were, according to the U.S. Census, 66177 non-African-American men between 18 and 29 with less than a high school education in Alabama, . . . , and 5 African American women over 65 with a college education in Wyoming.*

ethnicity, age, education, state, and number of people corresponding to each cell, as shown in Figure 14.4.

We then compute the expected response y^{pred}—the probability of supporting Bush for each cell. Assuming we have `n.sims` simulation draws after fitting the model in Bugs (see Chapter 16), we construct the following `n.sims` $\times$ 3264 matrix:

R code

```
L <- ncol (census)
y.pred <- array (NA, c(n.sims, L))
for (l in 1:L){
  y.pred[,l] <- invlogit(b.0 + b.female*census$female[l] +
    b.black*census$black[l] +
    b.female.black*census$female[l]*census$black[l] +
    a.age[,census$age[l]] + a.edu[,census$edu[l]] +
    a.age.edu[,census$age[l],census$edu[l]] + a.state[,census$state[l]])
}
```

For each state j, we are estimating the average response in the state,

$$y^{\text{pred}}_{\text{state}\,j} = \frac{\sum_{l \in j} N_l \theta_l}{\sum_{l \in j} N_l},$$

summing over the 64 demographic categories within the state. Here, we are using l as a general stratum indicator (not the same l used to index education categories in model (14.2); we are simply running out of "index"-type letters from the middle of the alphabet). The notation "$l \in j$" is shorthand for "category l represents a subset of state j." In R:

R code

```
y.pred.state <- array (NA, c(n.sims, n.state))
for (s in 1:n.sims){
  for (j in 1:n.state){
    ok <- census$state==j
    y.pred.state[s,j] <- sum(census$N[ok]*y.pred[s,ok])/sum(census$N[ok])
  }
}
```

We can then summarize these `n.sims` simulations to get a point prediction and a 50% interval for the proportion of adults in each state who supported Bush:

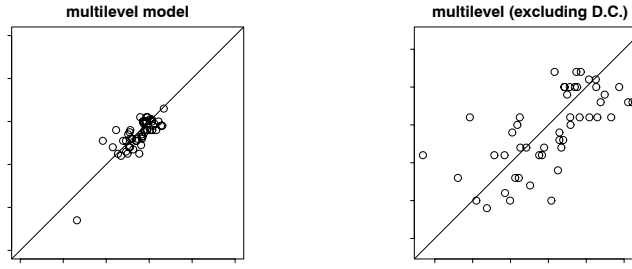

Figure 14.5 *For each state, the proportion of the two-party vote received by George Bush in 1988, plotted versus the support for Bush in the state, as estimated from a multilevel model applied to pre-election polls. The second plot excludes the District of Columbia in order to more clearly show the 50 states.*

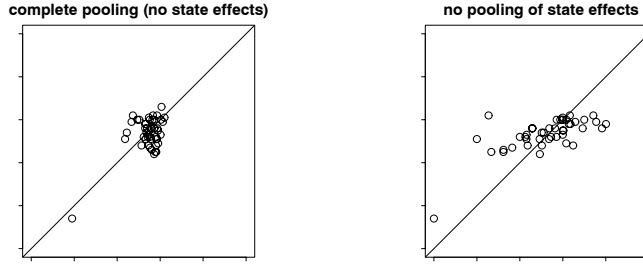

Figure 14.6 *For each state, Bush's vote in 1988 plotted versus his support in the polls, as estimated from (a) the complete-pooling model (using demographics alone with no state predictors), and (b) the no-pooling models (estimating each state separately). The two models correspond to* $\sigma_{\text{state}} = \sigma_{\text{region}} = 0$ *and* ∞*, respectively. Compare to Figure 14.5a, which shows results from the multilevel model (with* σ_{state} *and* σ_{region} *estimated from data).*

R code

```
state.pred <- array (NA, c(3,n.state))
for (j in 1:n.state){
  state.pred[,j] <- quantile (y.pred.state[,j], c(.25,.5,.75))
}
```

Comparing public opinion estimates to election outcomes

In this example, the estimates of the model come from opinion polls taken immediately before the election, and they can be externally validated by comparing to the actual election outcomes. We can thus treat this as a sort of "laboratory" for testing the accuracy of multilevel models and any other methods that might be used to estimate state-level opinions from national polls.

Figure 14.5 shows the actual election outcome for each state, compared to the model-based estimates of the proportion of Bush supporters. The fit is pretty good, with no strong systematic bias and an average absolute error of only 4.0%.

Comparison to simpler methods

By comparison, Figure 14.6 shows the predictive performance of the estimates based on complete pooling of states (estimating opinion solely based on demographics, thus setting $\alpha_j^{\text{state}} \equiv 0$ for all states) and no pooling (corresponding to completely separate estimates for each state, thus setting $\sigma_{\text{state}} = \sigma_{\text{region}} = \infty$). The complete-pooling model generally shrinks the state estimates too close toward the mean, whereas the no-pooling model does not shrink them enough. To make a numerical comparison, the average absolute error of the state estimates is 4.0% for the multilevel analysis, compared to 5.4% for complete pooling and 10.8% for no pooling.

14.2 Red states and blue states: what's the matter with Connecticut?

Throughout the twentieth century and even before, the Democratic Party in the United States has been viewed as representing the party of the lower classes and thus, by extension, the "average American." More recently, however, a different perspective has taken hold, in which the Democrats represent the elites rather than the masses. These patterns are complicated; on one hand, in recent U.S. presidential elections the Democrats have done best in the richer states of the Northeast and West (often colored blue in electoral maps) while the Republicans have dominated in the poorer "red states" in the South and between the coasts. On the other hand, using census and opinion poll data since 1952, we find that higher-income voters continue to support the Republicans in presidential elections.

We can understand these patterns, first by fitting a sequence of classical regressions and displaying estimates over time (as in Section 4.7), then by fitting some multilevel models:

- Aggregate, by state: to what extent do richer states favor the Democrats?
- Nationally, at the level of the individual voter: to what extent do richer voters support the Republicans?
- Individual voters within states: to what extent do richer voters support the Republicans, within any given state? In other words, how much does context matter?

We fit these models quickly with `lmer()` and then with Bugs, whose simulations we used to plot and understand the model. Here we describe the model and its estimate without presenting the steps of computation.

Classical regressions of state averages and individuals

Richer states now support the Democrats. We first present the comparison of red and blue states—more formally, regressions of Republican share of the two-party vote on state average per capita income (in tens of thousands of 1996 dollars). Figure 14.7a shows that, since the 1976 election, there has been a steady downward trend in the income coefficient over time. As time has gone on, richer states have increasingly favored the Democrats. For the past twenty years, the same patterns appear when fitting southern and non-southern states separately (Figure 14.7b,c).

Richer voters continue to support the Republicans overall. We fit a logistic regression of reported presidential vote preference ($y_i = 1$ for supporters of the Republican, 0 for the Democrats, and excluding respondents who preferred other candi-

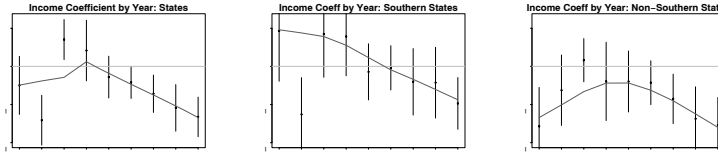

Figure 14.7 *(a) Regression predicting Republican vote share by average income in each state. The model was fit separately for each election year. Estimates and 95% error bars are shown. (b, c) Same model but fit separately to southern and non-southern states each year. Republicans do better in poor states than rich states, especially in recent years.*

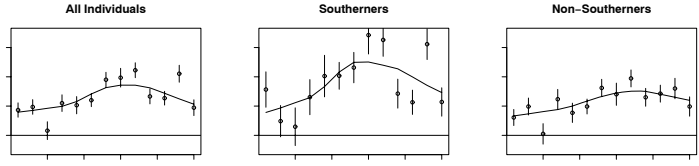

Figure 14.8 *Coefficients for income in logistic regressions of Republican vote, fit to National Election Studies data from each year. The positive coefficients indicate that higher-income voters have consistently supported the Republicans, a pattern that holds both within and outside the South.*

dates or expressed no opinion) on personal income,[1] fit separately to the National Election Study from each presidential election since 1952. Figure 14.8 shows that higher-income people have been consistently more likely to vote Republican. These patterns remain when ethnicity, sex, education, and age are added into the model: after controlling for these other individual-level predictors, the coefficient of income is still consistently positive.

A paradox? The conflicting patterns of Figures 14.7 and 14.8 have confused many political commentators. How can we understand the pattern of richer states supporting the Democrats, while richer voters support the Republicans? We shall use multilevel modeling to simultaneously study patterns within and between states.

Varying-intercept model of income and vote preference within states

We now focus on the 2000 presidential election using the National Annenberg Election Survey, which, with more than 100,000 respondents, allows accurate estimation of patterns within individual states. We fit a multilevel model that allows income to predict vote preference within each state, while also allowing systematic differences between states:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha_{j[i]} + \beta x_i), \quad \text{for } i = 1, \ldots, n, \tag{14.6}$$

[1] The National Election Study uses 1 = 0–16 percentile, 2 = 17–33 percentile, 3 = 34–67 percentile, 4 = 68–95 percentile, 5 = 96–100 percentile. We label these as $-2, -1, 0, 1, 2$, centering at zero (see Section 4.2) so that we can more easily interpret the intercept terms of regressions that include income as a predictor.

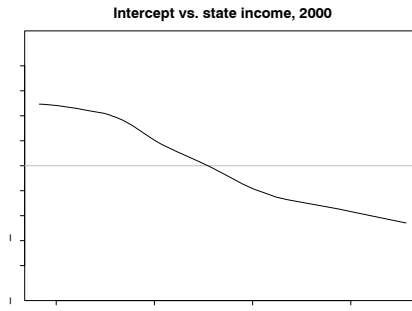

Figure 14.9 *Estimated state intercepts α_j in the varying-intercept logistic regression model (14.6)–(14.7) predicting Republican vote intention given individual income, plotted versus average state income. A nonparametric regression line fitted to the estimates is overlain for convenience.*

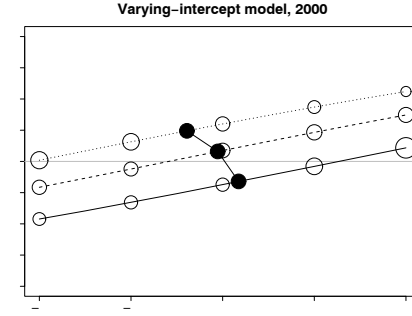

Figure 14.10 *The paradox is no paradox. From the multilevel logistic regression model for the 2000 election: probability of supporting Bush as a function of income category, for a a rich state (Connecticut), a medium state (Ohio), and a poor state (Mississippi). The open circles show the relative proportion (as compared to national averages) of households in each income category in each of the three states, and the solid circles show the average income level and estimated average support for Bush for each state. Within each state, richer people are more likely to vote Republican, but the states with higher income give more support to the Democrats.*

where $j[i]$ indexes the state (from 1 to 50) corresponding to respondent i, x_i is the person's household income (on the five-point scale), and n is the number of respondents in the poll.

We set up a state-level regression for the coefficients α_j, using the state average income level as a group-level predictor, which we label u_j:

$$\alpha_j \sim \text{N}(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2), \quad \text{for } j = 1, \ldots, 50. \tag{14.7}$$

Figure 14.9 shows the estimated state intercepts α_j, plotted versus average state income. There is a negative correlation between intercept and state income, which

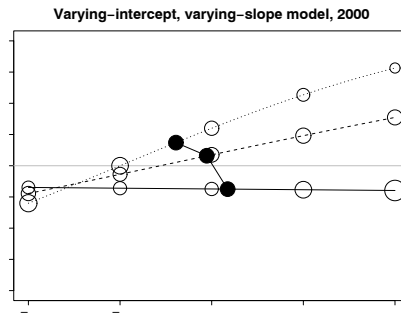

Figure 14.11 *From the multilevel logistic regression with varying intercepts and slopes for the 2000 election: probability of supporting Bush as a function of income category, for a a rich state (Connecticut), a medium state (Ohio), and a poor state (Mississippi). The open circles show the relative proportion (as compared to national averages) of households in each income category in each of the three states, and the solid circles show the average income level and estimated average support for Bush for each state. Income is a very strong predictor of vote preference in Mississippi, a weaker predictor in Ohio, and does not predict vote choice at all in Connecticut. See Figure 14.12 for estimated slopes in all 50 states, and compare to Figure 14.10, in which the state slopes are constrained to be equal.*

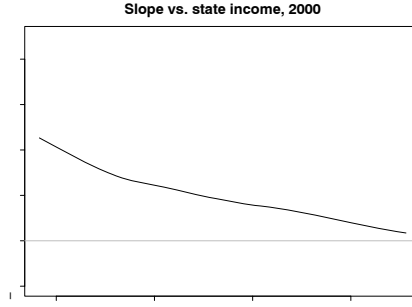

Figure 14.12 *Estimated coefficient for income within state plotted versus average state income, for the varying-intercept, varying-slope multilevel model (14.8)–(14.9) fit to the Annenberg survey data from 2000. A nonparametric regression line fitted to the estimates is overlain for convenience.*

tells us that, after adjusting for individual income, voters in richer states tend to support Democrats.

To understand the model as a whole, we display in Figure 14.10 the estimated logistic regression line, $\text{logit}^{-1}(\alpha_j + \beta x)$, for three states j: Connecticut (the richest state), Ohio (a state in the middle of the income distribution), and Mississippi (the poorest state). The graph shows a statistical resolution of the red-blue paradox. Within each state, income is positively correlated with Republican vote choice, but average income varies by state. For each of the three states in the plot, the open circles show the relative proportion of households in each income category (as

compared to national averages), and the solid circle shows the average income level and estimated average support for Bush in the state. The Bush-supporting states have more lower-income people, and as a result there is a negative correlation between average state income and state support for Bush, even amid the positive slope for each state. The poor people in "red" (Republican-leaning) states tend to be Democrats; the rich people in "blue" (Democratic-leaning) states tend to be Republicans. Income matters; also geography matters. Individual income is a positive predictor, and state average income is a negative predictor, of Republican presidential vote support.

Varying-intercept, varying-slope model

As Figure 14.10 shows, income and state are both predictive of vote preference. It is thus natural to consider their interaction, which in a multilevel context is a varying-intercept, varying-slope model:

$$\Pr(y_i=1) = \text{logit}^{-1}(\alpha_{j[i]} + \beta_{j[i]}x_i), \quad \text{for } i = 1, \ldots, n, \tag{14.8}$$

where, as in (14.6), x_i is respondent i's income (on the -2 to $+2$ scale). The state-level intercepts and slopes that are themselves modeled given average state incomes u_j:

$$\begin{aligned} \alpha_j &= \gamma_0^\alpha + \gamma_1^\alpha u_j + \epsilon_j^\alpha, \quad \text{for } j = 1, \ldots, 50 \\ \beta_j &= \gamma_0^\beta + \gamma_1^\beta u_j + \epsilon_j^\beta, \quad \text{for } j = 1, \ldots, 50, \end{aligned} \tag{14.9}$$

with errors $\epsilon_j^\alpha, \epsilon_j^\beta$ having mean 0, variances $\sigma_\alpha^2, \sigma_\beta^2$, and correlation ρ, all estimated from data. By including average income as a state-level predictor, we are not requiring the intercepts and slopes to vary linearly with income—the error terns ϵ_j allow for deviation from the model—but rather are allowing the model to find such linear relations to the extent they are supported by the data.

From this new model, we indeed find strong variation among states in the role of income in predicting vote preferences. Figure 14.11 recreates Figure 14.10 with the estimated varying intercepts and slopes. As before, we see generally positive slopes within states and a negative slope between states. What is new, though, is a systematic pattern of the within-state slopes, with the steepest slope in the poorest state—Mississippi—and the shallowest slope in the richest state—Connecticut.

Figure 14.12 shows the estimated slopes for all 50 states and reveals a clear pattern, with high coefficients—steep slopes—in poor states and low coefficients in rich states. Income matters more in "red America" than in "blue America." The varying-intercept, varying-slope multilevel model has been a direct approach for us to discover these patterns.

14.3 Item-response and ideal-point models

We could have introduced these in Chapter 6 in the context of classical generalized linear models, but item-response and ideal-point models are always applied to data with multilevel structure, typically non-nested, for example with measurements associated with persons and test items, or judges and cases. As with the example of the previous section, we present the models here, deferring computation until the presentation of Bugs in Part 2B of this book.

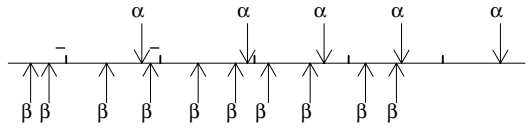

Figure 14.13 *Illustration of the logistic item-response (Rasch) model,* $\Pr(y_i = 1) = logit^{-1}(\alpha_{j[i]} - \beta_{k[i]})$*, for an example with 5 persons* j *(with abilities* α_j*) and 10 items* k *(with difficulties* β_k*). If your ability* α *is greater than the difficulty* β *of an item, then you have a better-than-even chance of getting that item correct. This graph also illustrates the nonidentifiability in the model: the probabilities depend only on the* relative *positions of the ability and difficulty parameters; thus, a constant could be added to all the* α_j*'s and all the* β_k*'s, and the model would be unchanged. One way to resolve this nonidentifiability is to constrain the* α_j*'s to have mean 0. Another solution is to give the* α_j*'s a distribution with mean fixed at 0.*

The basic model with ability and difficulty parameters

A standard model for success or failure in testing situations is the logistic item-response model, also called the Rasch model. Suppose J persons are given a test with K items, with $y_{jk} = 1$ if the response is correct. Then the logistic model can be written as

$$\Pr(y_{jk}=1) = \text{logit}^{-1}(\alpha_j - \beta_k), \tag{14.10}$$

with parameters:

- α_j: the *ability* of person j
- β_k: the *difficulty* of item k.

In general, not every person is given every item, so it is convenient to index the individual responses as $i = 1, \ldots, n$, with each response i associated with a person $j[i]$ and item $k[i]$. Thus model (14.10) becomes

$$\Pr(y_i=1) = \text{logit}^{-1}(\alpha_{j[i]} - \beta_{k[i]}). \tag{14.11}$$

Figure 14.13 illustrates the model as it might be estimated for 5 persons with abilities α_j, and 10 items with difficulties β_k. In this particular example, questions 5, 3, and 8 are easy questions (relative to the abilities of the persons in the study), and all persons except person 2 are expected to answer more than half the items correctly. More precise probabilities can be calculated using the logistic distribution: for example, α_2 is 2.4 higher than β_5, so the probability that person 2 correctly answers item 5 is $\text{logit}^{-1}(2.4) = 0.92$, or 92%.

Identifiability problems

This model is not identified, whether written as (14.10) or as (14.11), because a constant can be added to all the abilities α_j and all the difficulties β_k, and the predictions of the model will not change. The probabilities depend only on the *relative* positions of the ability and difficulty parameters. For example, in Figure 14.13, the scale could go from -104 to -96 rather than -4 to 4, and the model would be unchanged—a difference of 1 on the original scale is still a difference of 1 on the shifted scale.

From the standpoint of classical logistic regression, this nonidentifiability is a simple case of collinearity and can be resolved by constraining the estimated parameters in some way: for example, setting $\alpha_1 = 0$ (that is, using person 1 as a

"baseline"), setting $\beta_1 = 0$ (so that a particular item is the comparison point), constraining the α_j's to sum to 0, or constraining the β_j's to sum to 0. In a multilevel model, such constraints are unnecessary, as we discuss next.

Multilevel model

The natural multilevel model for (14.11) assigns normal distributions to the ability and difficulty parameters:

$$\begin{aligned} \alpha_j &\sim \text{N}(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J \\ \beta_k &\sim \text{N}(\mu_\beta, \sigma_\beta^2), \text{ for } k = 1, \ldots, K. \end{aligned}$$

This model is nonidentified for the reasons discussed above: now it is μ_α and μ_β that are not identified, because a constant can be added to each without changing the predictions. The simplest way to identify the multilevel model is set μ_α to 0, or to set μ_β to 0 (but not both).

As usual, we can add group-level predictors. In this case, the "groups" are the persons and items:

$$\begin{aligned} \alpha_j &\sim \text{N}(X_j^\alpha \gamma_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J \\ \beta_k &\sim \text{N}(X_j^\beta, \sigma_\beta^2), \text{ for } k = 1, \ldots, K. \end{aligned}$$

In an educational testing example, the person-level predictors X^α could include age, sex, and previous test scores, and the item-level predictors X^β could include a prior measure of item difficulty (perhaps the average score for that item from a previous administration of the test).

Defining the model using redundant parameters

Another way to identify the model is by allowing the parameters α and β to "float" and then defining new quantities that are well identified. The new quantities can be defined, for example, by rescaling based on the mean of the α_j's:

$$\begin{aligned} \alpha_j^{\text{adj}} &= \alpha_j - \bar{\alpha}, \text{ for } j = 1, \ldots, J \\ \beta_k^{\text{adj}} &= \beta_k - \bar{\alpha}, \text{ for } k = 1, \ldots, K. \end{aligned} \tag{14.12}$$

The new ability parameters α_j^{adj} and difficulty parameters β_k^{adj} are well defined, and they work in place of α and β in the original model:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha_{j[i]}^{\text{adj}} - \beta_{k[i]}^{\text{adj}}).$$

This holds because we subtracted the same constant from the α's and β's in (14.12). For example, it would *not* work to subtract $\bar{\alpha}$ from the α_j's and $\bar{\beta}$ from the β_k's because then we would lose our ability to distinguish the position of the parameters relative to each other.

Adding a discrimination parameter

The item-response model can be generalized by allowing the slope of the logistic regression to vary by item:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\gamma_{k[i]}(\alpha_{j[i]} - \beta_{k[i]})). \tag{14.13}$$

In this new model, γ_k is called the *discrimination* of item k: if $\gamma_k = 0$, then the item does not "discriminate" at all ($\Pr(y_i = 1) = 0.5$ for any person), whereas high

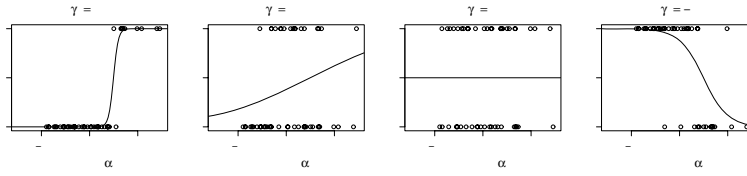

Figure 14.14 *Curves and simulated data from the logistic item-response (Rasch) model for items k with "difficulty" parameter $\beta_k = 1$ and high, low, zero, and negative "discrimination" parameters γ_k.*

values of γ_k correspond to strong relation between ability and the probability of getting a correct response. Figure 14.14 illustrates.

In educational testing, it is generally desirable for items k to have high values of γ_k, because the responses to these items can better "discriminate" between high and low abilities (see the left plot in Figure 14.14). The ideal test would have several items, each with high γ_k, and with difficulties β_k that span the range of the abilities of the persons being tested. Items with γ_k near zero do not do a good job at discriminating between abilities (see the center two plots in Figure 14.14), and negative values of γ_k correspond to items where low-ability persons do better. Such items typically represent mistakes in the construction of the test.

Including the discrimination parameter creates additional identifiability problems which we will discuss in the context of an example in the next section.

An ideal-point model for Supreme Court voting

Ideal-point modeling is an application of item-response models to a setting where what is being measured is not "ability" of individuals and "difficulty" of items, but rather positions of individuals and items on some scale of values.

We illustrate with a study of voting records of U.S. Supreme Court justices, using all the Court's decisions since 1954. Each vote i is associated with a justice $j[i]$ and a case $k[i]$, with an outcome y_i that equals 1 if the justice voted "yes" on the case and 0 if "no." In this particular example, the votes have been coded so that a "yes" response ($y_i = 1$) is intended to correspond to the politically "conservative" outcome, with "no" ($y_i = 0$) corresponding to a "liberal" vote.

As with the item-response models discussed above, the data are modeled with a logistic regression, with the probability of voting "yes" depending on the "ideal point" α_j for each justice, the "position" β_k for each case, and a "discrimination parameter" γ_k for each case, following the three-parameter logistic model (14.13).

The positions on this scale (equivalent to the α's and β's on Figure 14.14) represent whatever dimension is best able to explain the voting patterns. For the Supreme Court, we represent it as an ideological dimension, with liberal justices having positions on the left side of the scale (negative α_j's) and conservatives being on the right side (positive α_j's).

For any given justice j and case k, the difference between α_j and β_k indicates the relative positions of the justice and the case—if a justice's ideal point is near a case's position, then the case could go either way, but if the ideal point is far from the position, then the justice's vote is highly predictable. The discrimination parameter γ_k captures the importance of the positioning in determining the justices' votes: if $\gamma_k = 0$, the votes on case k are purely random; and if γ_k is very large (in

absolute value), then the relative positioning of justice and case wholly determines the outcome. Changing the sign of γ changes which justices are expected to vote yes and which to vote no.

Model (14.13) has two indeterminacies: an additive *aliasing* in α and β (that is, a situation in which values of α and β can be changed while keeping the model's predictions unchanged), and a multiplicative aliasing in all three parameters. The additive aliasing occurs because a constant can be added to all the α's and all the β's, leaving the model predictions (and thus the likelihood) unchanged. The multiplicative aliasing arises when multiplying the γ's by a constant and dividing the α's and β's by that same constant. We can resolve both these indeterminacies by constraining the α_j's to have mean 0 and standard deviation 1 or, in a multilevel context, by giving the α_j a $\text{N}(0, 1)$ distribution. In contrast the parameters β and γ are unconstrained (or, in a multilevel context, have $\text{N}(\mu_\beta, \sigma_\beta^2)$ and $\text{N}(\mu_\gamma, \sigma_\gamma^2)$ distributions whose means and variances are estimated from the data, as part of a multilevel model).

Even after constraining the distribution of the position parameters α_j, one indeterminacy remains in model (14.13): a reflection invariance associated with multiplying all the γ_k's, α_j's, and β_k's by -1. If no additional constraints are assigned to this model, this aliasing will cause a bimodal likelihood and posterior distribution. It is desirable to select just one of these modes for our inferences. (Among other problems, if we include both modes, then each parameter will have two maximum likelihood estimates and a posterior mean of 0.)

We first briefly discuss two simple and natural ways of resolving the aliasing. The first approach is to constrain the γ's to all have positive signs. This might seem to make sense, since the outcomes have been precoded so that positive y_i's correspond to conservative votes. However, we do not use this approach because it relies too strongly on the precoding, which, even if it is generally reasonable, is not perfect. We would prefer to estimate the ideological direction of each vote from the data and then compare to the precoding to check that the model makes sense (and to explore any differences found between the estimates and the precoding).

A second approach to resolving the aliasing is to choose one of the α's, β's, or γ's, and restrict its sign or choose two and constrain their relative position. For example, we could constrain α_j to be negative for the extremely liberal William Douglas, or constrain α_j to be positive for the extremely conservative Antonin Scalia. Or, we could constrain Douglas's α_j to be less than Scalia's α_j.

Only a single constraint is necessary to resolve the two modes; if possible, however, it should be a clear-cut division. One can imagine a general procedure that would be able to find such divisions based on the data, but in practice it is simpler to constrain using prior information such as the identification of extremely liberal and conservative judges in this example. (Not all choices of constraints would work. For example, if we were to constrain $\alpha_j > 0$ for a merely moderately conservative judge such as Sandra Day O'Connor, this could split the likelihood surface across both modes, rather than cleanly selecting a single mode.)

The alternative approach we actually use in this example is to encode the additional information in the form of a group-level regression predictor, whose coefficient we constrain to be positive. Various case-level and justice-level predictors can be added to model (14.13), but the simplest is an indicator that equals 1 for Scalia, -1 for Douglas, and 0 for all other justices. We set up a multilevel model for the justices' ideal points,

$$\alpha_j = \delta x_j + \text{error}_j \tag{14.14}$$

where x_j is this Scalia/Douglas predictor. Constraining the regression coefficient

$\delta > 0$ identifies the model (by aligning the positive direction with the difference between these two extreme justices) but in a flexible way that allows us to estimate our full model.

Two-dimensional item-response or ideal-point models

In a two-dimensional item-response model, the task of getting an item correct requires a combination of two "skills," which can be represented for each person j as a two-dimensional "ability" vector $(\alpha_j^{(1)}, \alpha_j^{(2)})$. (For example, on a high school general aptitude test, the two dimensions might correspond to verbal and mathematical ability.) The two-dimensional "difficulty" parameter $(\beta_k^{(1)}, \beta_k^{(2)})$ represents the thresholds required to perform well on the task, and the discrimination parameters $\gamma_k^{(1)}, \gamma_k^{(2)}$ indicate the relevance of each of the two skills to task k.

Success on the two skills can be combined in a variety of ways. For example, in a "conjunctive" model, both skills are required to perform the task correctly; thus,

$$\begin{aligned}\text{conjunctive model:}\quad \Pr(y_i = 1) &= \text{logit}^{-1}\left[\gamma_{k[i]}^{(1)}\left(\alpha_{j[i]}^{(1)} - \beta_{k[i]}^{(1)}\right)\right] \\ &\quad \times \text{logit}^{-1}\left[\gamma_{k[i]}^{(2)}\left(\alpha_{j[i]}^{(2)} - \beta_{k[i]}^{(2)}\right)\right].\end{aligned}$$

In a "disjunctive" model, either skill is sufficient to perform the task:

$$\begin{aligned}\text{disjunctive model:}\quad 1 - \Pr(y_i = 1) &= \left(1 - \text{logit}^{-1}\left[\gamma_{k[i]}^{(1)}\left(\alpha_{j[i]}^{(1)} - \beta_{k[i]}^{(1)}\right)\right]\right) \\ &\quad \times \left(1 - \text{logit}^{-1}\left[\gamma_{k[i]}^{(2)}\left(\alpha_{j[i]}^{(2)} - \beta_{k[i]}^{(2)}\right)\right]\right).\end{aligned}$$

Perhaps the most straightforward model is additive on the logistic scale:

$$\begin{aligned}\text{additive model:}\quad \Pr(y_i = 1) &= \text{logit}^{-1}\left[\gamma_{k[i]}^{(1)}\left(\alpha_{j[i]}^{(1)} - \beta_{k[i]}^{(1)}\right)\right. \\ &\quad \left. + \gamma_{k[i]}^{(2)}\left(\alpha_{j[i]}^{(2)} - \beta_{k[i]}^{(2)}\right)\right].\end{aligned}$$

In the "ideal-point" formulation of these models, α_j represents the ideal point of justice j in two dimensions (for example, a left-right dimension for economic issues, and an authoritarian-libertarian dimension on social issues), β_k is the indifference point for case k in these dimensions, the signs of the two components of γ_k give the direction of a Yes vote in terms of the two issues, and the absolute values of $\gamma_k^{(1)}, \gamma_k^{(2)}$ indicate the importance of each issue in determining the vote.

Other generalizations

As formulated so far, the probabilities in the item-response and ideal-point models range from 0 to 1 and are symmetric about 0.5 (see Figure 14.14). Real data do not necessarily look like this. One simple way to generalize the model is to limit the probabilities to a fixed range:

$$\Pr(y_i = 1) = \pi_1 + (1 - \pi_0 - \pi_1)\text{logit}^{-1}[\gamma_{k[i]}(\alpha_{j[i]} - \beta_{k[i]})].$$

In this model, every person has an immediate probability of π_1 of success and π_0 of failure, with the logistic regression model applying to the remaining $(1 - \pi_0 - \pi_1)$ of outcomes. For example, suppose we are modeling responses to a multiple-choice exam, and $\pi_1 = 0.25$ and $\pi_0 = 0.05$. We could interpret this as a 25% chance of getting an item correct by guessing, along with a 5% chance of getting an item wrong by a careless mistake.

Another way to generalize item response and ideal point models is to go beyond the logistic distribution, for example using a robit model as described in Section 6.6 that allows for occasional mispredictions.

14.4 Non-nested overdispersed model for death sentence reversals

So far in this chapter we have presented logistic regression for binary data points y_i that can equal 0 or 1. The model can also be used for proportions, in which each data point y_i equals the number of "successes" out of n_i chances. For example, Section 6.3 describes data on death penalty reversals, in which i indexes state-years (for example, Alabama in 1983), n_i is the number of death sentences given out in that particular state in that particular year, and y_i is the number of these death sentences that were reversed by a higher court. We now describe how we added multilevel structure to this model.

Non-nested model for state and year coefficients

The death penalty model had several predictors in X, including measures of the frequency that the death sentence was imposed, the backlog of capital cases in the appeals courts, the level of political pressure on judges, and other variables at the state-year level.

In addition, we included indicators for the years from 1973 to 1995 and the 34 states (all of those in this time span that had death penalty laws). The regression model with all these predictors can be written as

$$\begin{aligned} y_i &\sim \text{Bin}(n_i, p_i) \\ p_i &= \text{logit}^{-1}(X_i\beta + \alpha_{j[i]} + \gamma_{t[i]}), \end{aligned} \tag{14.15}$$

where j indexes states and t indexes years. We complete the multilevel model with distributions for the state and year coefficients,

$$\begin{aligned} \alpha_j &\sim \text{N}(0, \sigma_\alpha^2) \\ \gamma_t &\sim \text{N}(a + bt,\, \sigma_\gamma^2). \end{aligned}$$

The coefficients for year include a linear time trend to capture the overall increase in reversal rates during the period under study. The model for the γ_t's also includes an intercept, and so we do not need to include a constant term in the model for the α_j's or in the matrix X of individual-level predictors in (14.15).

In this particular example, we are not particularly interested in the coefficients for individual states or years; rather, we want to include these sources of variability into the model in order to get appropriate uncertainty estimates for the coefficients of interest, β.

Multilevel overdispersed binomial regression

Testing for overdispersion. Model (14.15) is inappropriate for the death penalty data because the data are overdispersed, as discussed in Section 6.3. To measure the overdispersion, we compute the standardized residuals, $z_i = (y_i - p_i)/\sqrt{p_i(1-p_i)/n_i}$ with p_i as defined in (14.15). Under the binomial model, the residuals should have mean 0 and standard deviation 1, and so $\sum_i z_i^2$ should look like a random draw from a χ^2 distribution with degrees of freedom equal to 520 (the number of state-years in the data).

Testing for overdispersion in a classical binomial regression is described in Section

6.3, where the z_i's are computed based on estimated probabilities $\hat{p}_i$, and $\sum_i z_i^2$ is compared to a χ^2 distribution with degrees of freedom adjusted for the number of coefficients estimated in the model.

Beta-binomial model. There are two natural overdispersed generalizations of the multilevel binomial regression (14.15). The first approach uses the beta-binomial distribution:

$$y_i \sim \text{beta-binomial}(n_i, p_i, \omega),$$

where $\omega \geq 1$ is the overdispersion parameter (and the model with $\omega = 1$ reduces to the binomial).

Binomial-normal model. The other direct way to construct an overdispersed binomial distribution is to add normal errors on the logistic scale, keeping the binomial model but adding a data-level error ξ_i to the linear predictor in (14.15):

$$p_i = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]} + \gamma_{t[i]} + \xi_i),$$

with these errors having their own normal distribution:

$$\xi_i \sim \text{N}(0, \sigma_\xi^2).$$

The resulting model reduces to the binomial when $\sigma_\xi = 0$; otherwise it is overdispersed.

With moderate sample sizes, it is typically difficult to distinguish between the beta-binomial and binomial-normal models, and the choice between them is one of convenience. The beta-binomial model adds only one new parameter and so can be easier to fit; however, the binomial-normal model has the advantage that the new error term ξ_i is on the same scale as the group-level predictors, α_j and γ_t, which can make the fitted model easier to understand.

14.5 Bibliographic note

Multilevel logistic regression has a long history in the statistical and applied literature which we do not attempt to trace here: the basic ideas are the same as in multilevel linear models (see references in Sections 12.10 and 13.8) but with complications arising from the discreteness of the data and the nonlinearity of some of the computational steps.

The example of state-level opinions from national polls comes from Gelman and Little (1997) and Park, Gelman, and Bafumi (2004). The analysis of income and voting comes from Gelman, Shor, et al. (2005); see also Wright (1989), Ansolabehere, Rodden, and Snyder (2005), and McCarty, Poole, and Rosenthal (2005) for related work. Wainer (2002) discusses "B-K" plots (named after), which similar to Figure 14.10, which simultaneously displays patterns within and between groups, is related to the "B-K plot" (discussed by Wainer, 2002, and named after Baker and Kramer, 2001).

The multilevel framework for item-response and ideal-point models appears in Bafumi, Gelman, and Park (2005). See Lord and Novick (1968) and van der Linden and Hambleton (1997) for more on item-response models, and Poole and Rosenthal (1997), Jackman (2001), and Martin and Quinn (2002a) for more on ideal-point models. Loken (2004) discusses identifiability problems in models with aliasing.

The death sentencing example comes from Gelman, Liebman, et al. (2004). See Donohue and Wolfers (2006) for an overview of some of the research literature on death sentencing.

14.6 Exercises

1. The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, political ideology, and state.

   (a) Fit a logistic regression predicting support for Bush given all these inputs except state. Consider how to include these as regression predictors and also consider possible interactions.

   (b) Now formulate a model predicting support for Bush given the same inputs but allowing the intercept to vary over state. Fit using `lmer()` and discuss your results.

   (c) Create graphs of the probability of choosing Bush given the linear predictor associated with your model separately for each of eight states as in Figure 14.2.

2. The well-switching data described in Section 5.4 are in the folder `arsenic`.

   (a) Formulate a multilevel logistic regression model predicting the probability of switching using log distance (to nearest safe well) and arsenic level and allowing intercepts to vary across villages. Fit this model using `lmer()` and discuss the results.

   (b) Extend the model in (b) to allow the coefficient on arsenic to vary across village, as well. Fit this model using `lmer()` and discuss the results.

   (c) Create graphs of the probability of switching wells as a function of arsenic level for eight of the villages.

   (d) Compare the fit of the models in (a) and (b).

3. Three-level logistic regression: the folder `rodents` contains data on rodents in a sample of New York City apartments.

   (a) Build a varying intercept logistic regression model (varying over buildings) to predict the presence of rodents (the variable `rodent2` in the dataset) given indicators for the ethnic groups (`race`) as well as other potentially relevant predictors describing the apartment and building. Fit this model using `lmer()` and interpret the coefficients at both levels.

   (b) Now extend the model in (b) to allow variation across buildings within community district and then across community districts. Also include predictors describing the community districts. Fit this model using `lmer()` and interpret the coefficients at all levels.

   (c) Compare the fit of the models in (a) and (b).

4. Item-response model: the folder `exam` contains data on students' success or failure (item correct or incorrect) on a number of test items. Write the notation for an item-response model for the ability of each student and level of difficulty of each item.

5. Multilevel logistic regression with non-nested groupings: the folder `speed.dating` contains data from an experiment on a few hundred students that randomly assigned each participant to 10 short dates with participants of the opposite sex (Fisman et al., 2006). For each date, each person recorded several subjective numerical ratings of the other person (attractiveness, compatibility, and some

other characteristics) and also wrote down whether he or she would like to meet the other person again. Label

$$y_{ij} = \begin{cases} 1 & \text{if person } i \text{ is interested in seeing person } j \text{ again} \\ 0 & \text{otherwise} \end{cases}$$

and $r_{ij1}, \ldots, r_{ij6}$ as person i's numerical ratings of person j on the dimensions of attractiveness, compatibility, and so forth.

(a) Fit a classical logistic regression predicting $\Pr(y_{ij} = 1)$ given person i's 6 ratings of person j. Discuss the importance of attractiveness, compatibility, and so forth in this predictive model.

(b) Expand this model to allow varying intercepts for the persons making the evaluation; that is, some people are more likely than others to want to meet someone again. Discuss the fitted model.

(c) Expand further to allow varying intercepts for the persons being rated. Discuss the fitted model.

6. Varying-intercept, varying-slope logistic regression: continuing with the speed-dating example from the previous exercise, you will now fit some models that allow the coefficients for attractiveness, compatibility, and the other attributes to vary by person.

(a) Fit a no-pooling model: for each person i, fit a logistic regression to the data y_{ij} for the 10 persons j whom he or she rated, using as predictors the 6 ratings $r_{ij1}, \ldots, r_{ij6}$. (Hint: with 10 data points and 6 predictors, this model is difficult to fit. You will need to simplify it in some way to get reasonable fits.)

(b) Fit a multilevel model, allowing the intercept and the coefficients for the 6 ratings to vary by the rater i.

(c) Compare the inferences from the multilevel model in (b) to the no-pooling model in (a) and the complete-pooling model from part (a) of the previous exercise.