IDS 702: Module 4.5

Multilevel/hierarchical logistic models

Dr. Olanrewaju Michael Akande

Multilevel linear models

- The same idea and approach used to build multilevel models for normal data can be used to build multilevel logistic (and probit) models for binary outcomes.

- Recall that for a varying-intercepts linear model with one individual-level predictor, we have

$$\begin{aligned}
y_{ij} &= \beta_0 + \gamma_{0j} + \beta_1 x_{1ij} + \epsilon_{ij}; \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J; \\
\epsilon_{ij} &\sim N(0, \sigma^2); \\
\gamma_{0j} &\sim N(0, \tau_0^2)
\end{aligned}$$

  where x_{1ij} can be replaced with x_{1j} for a group-level predictor.

- This model is a compromise between complete pooling across groups of a grouping variable, such as counties in the radon example for last class (that is, same intercept for each county), and no pooling (estimating a separate intercept for each county without borrowing information).

- The degree of pooling is determined by the amount of information within and between groups.

IDS 702

Multilevel logistic models

- We can use the same idea to build a varying-intercepts logistic model.
- That is,

$$
\begin{aligned}
y_{ij}|x_{ij} &\sim \text{Bernoulli}(\pi_{ij}); \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J; \\
\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + \gamma_{0j} + \beta_1 x_{1ij}; \\
\gamma_{0j} &\sim N(0, \sigma_0^2)
\end{aligned}
$$

  where once again, x_{1ij} is an individual-level predictor which can be replaced with x_{1j} for a group-level predictor.

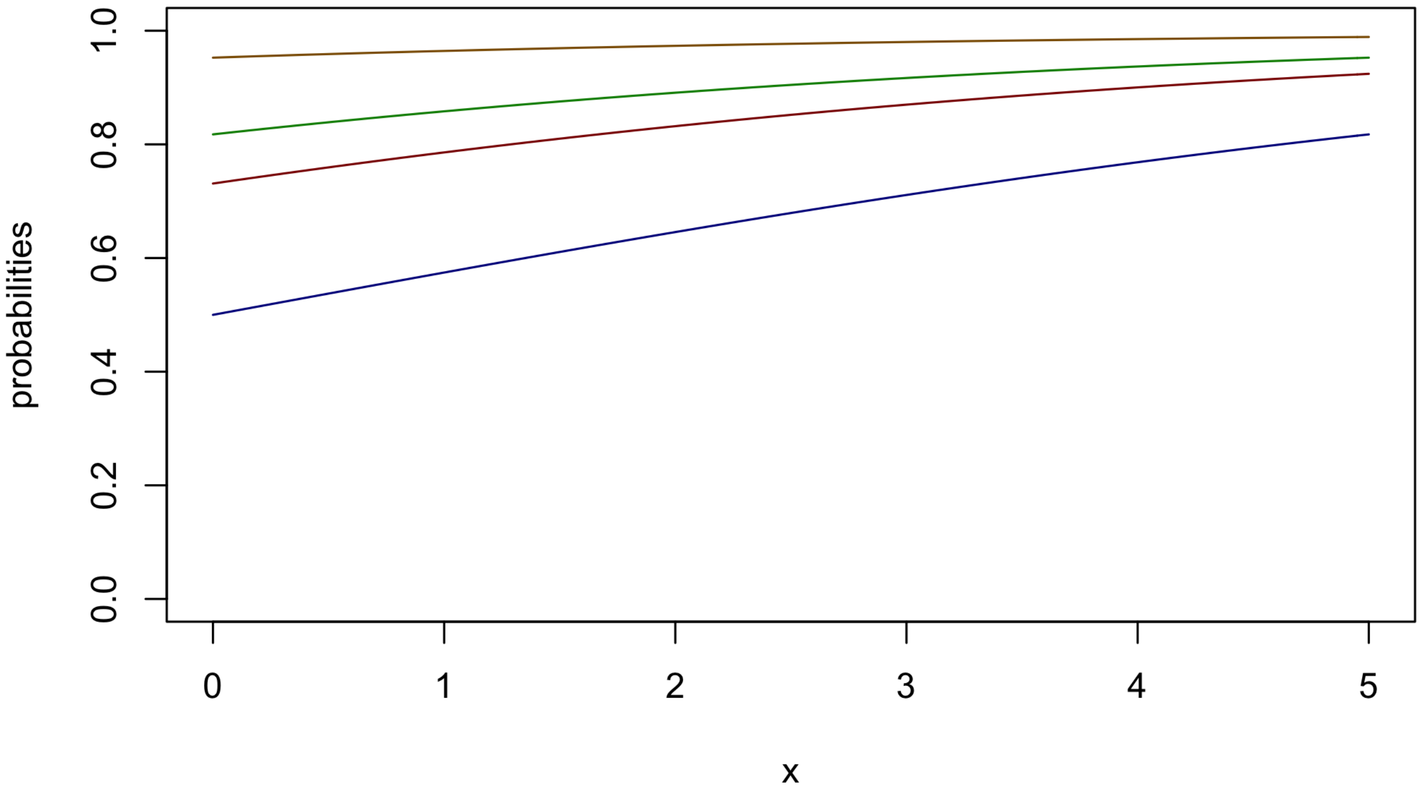
- The Gelman and Hill book uses the following notation instead

$$
\begin{aligned}
y_i|x_i &\sim \text{Bernoulli}(\pi_i); \quad i = 1, \ldots, n; \quad j = 1, \ldots, J; \\
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \gamma_{0j[i]} + \beta_1 x_{i1}; \\
\gamma_{0j} &\sim N(0, \sigma_0^2).
\end{aligned}
$$

- I will use this notation in this module and the next.

Varying-intercepts logistic model

Inverse logit functions for varying-intercepts logistic models.



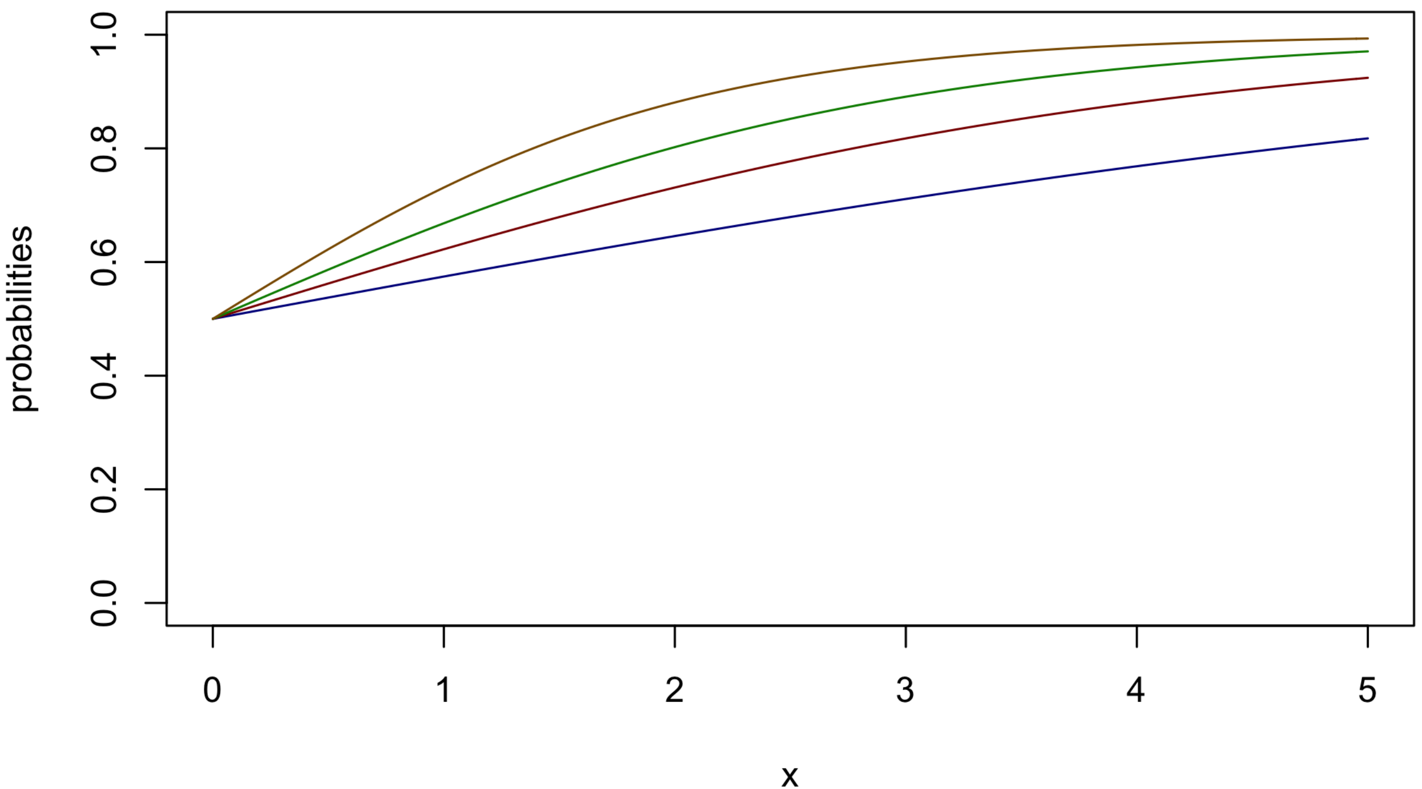
IDS 702

Multilevel logistic models

- It is easy to extend this model to allow for varying-slopes or both varying-intercepts and varying-slopes just like we had for multilevel linear models.
- The interpretations of the fixed effect(s) in multilevel logistic models follow directly from what we had for the standard logistic models, that is, log-odds, odds and odds ratios.
- The only difference now is the hierarchy in our data which allows us to borrow information across groups.
- One way to think about this is that we expect odds and odd-ratios to be more similar for observations within the same group, but we allow for some similarity across groups via partial pooling.

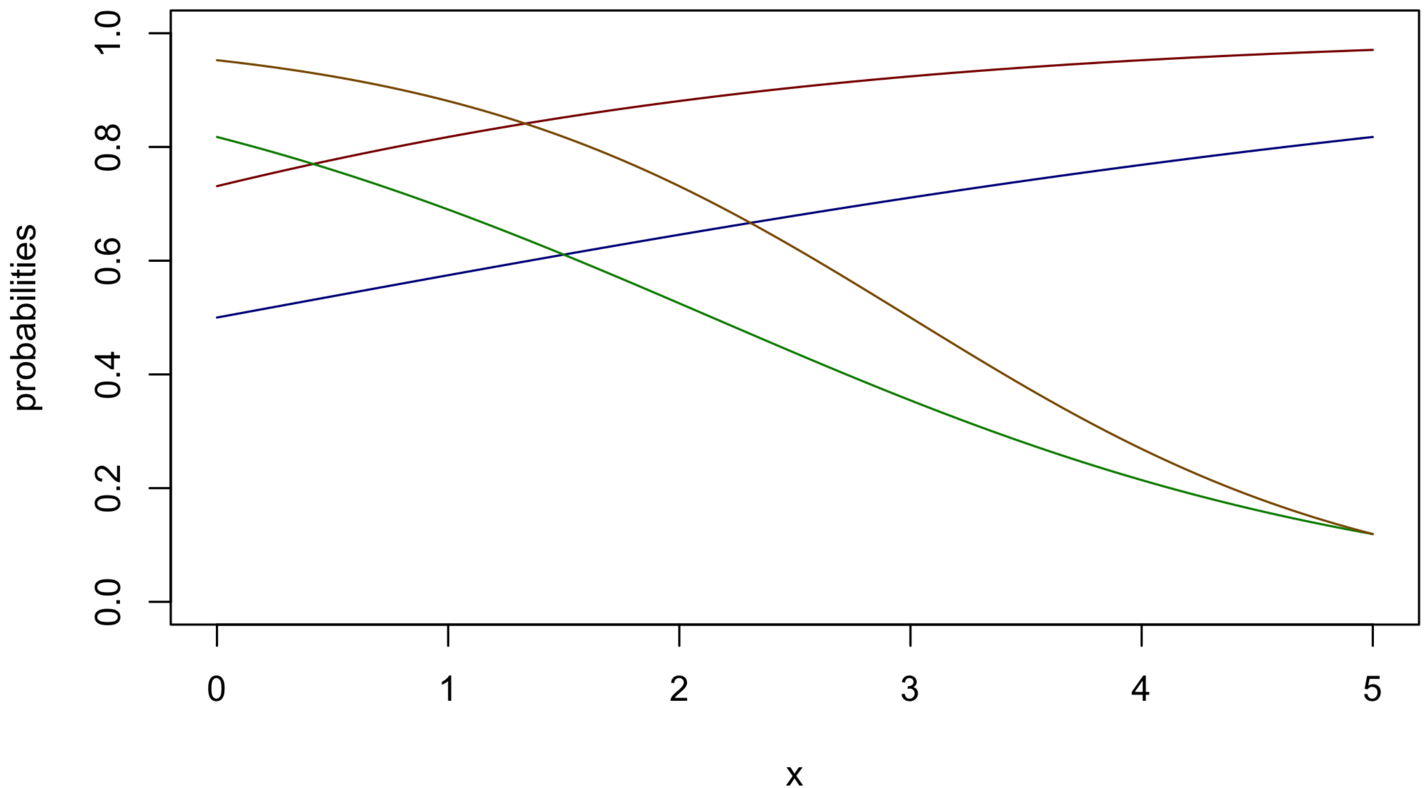
Varying-slopes logistic model

Inverse logit functions for varying-slopes logistic models.



Varying-intercepts, varying-slopes logistic model

Inverse logit functions for varying-intercepts, varying-slopes logistic models.



IDS 702

1988 ELECTIONS ANALYSIS

- To illustrate how to fit and interpret the results of multilevel logistic models, we will use a sample data on election polls.
- National opinion polls are conducted by a variety of organizations (e.g., media, polling organizations, campaigns) leading up to elections.
- While many of the best opinion polls are conducted at a national level, it can also be often interesting to estimate voting opinions and preferences at the state or even local level.
- Well-designed polls are generally based on national random samples with corrections for nonresponse based on a variety of demographic factors (e.g., sex, ethnicity, race, age, education).
- The data is from CBS News surveys conducted during the week before the 1988 election.
- Respondents were asked about their preferences for either the Republican candidate (Bush Sr.) or the Democratic candidate (Dukakis).

1988 ELECTIONS ANALYSIS

The dataset includes 2193 observations from one of eight surveys (the most recent CBS News survey right before the election) in the original full data.

Variable	Description
org	cbsnyt = CBS/NYT
bush	1 = preference for Bush Sr., 0 = otherwise
state	1-51: 50 states including DC (number 9)
edu	education: 1=No HS, 2=HS, 3=Some College, 4=College Grad
age	1=18-29, 2=30-44, 3=45-64, 4=65+
female	1=female, 0=male
black	1=black, 0=otherwise
region	1=NE, 2=S, 3=N, 4=W, 5=DC
v_prev	average Republican vote share in the three previous elections (adjusted for home-state and home-region effects in the previous elections)

Given that the data has a natural multilevel structure (through `state` and `region`), it makes sense to explore multilevel models for this data.

We will do just that in the next module.

IDS 702

1988 ELECTIONS ANALYSIS

- Both voting turnout and preferences often depend on a complex combination of demographic factors.
- In our example dataset, we have demographic factors such as biological sex, race, age, education, which we may all want to look at by state, resulting in $2 \times 2 \times 4 \times 4 \times 51 = 3264$ potential categories of respondents.
- We may even want to control for `region`, adding to the number of categories.
- Clearly, without a very large survey (most political survey poll around 1000 people), we will need to make assumptions in order to even obtain estimates in each category.
- We usually cannot include all interactions; we should therefore select those to explore (through EDA and background knowledge).
- The data is in the file `polls_subset.txt` on Sakai.

1988 ELECTIONS ANALYSIS

```r
###### Load the data
polls_subset <- read.table("data/polls_subset.txt",header=TRUE)
str(polls_subset)
```

```
## 'data.frame':    2193 obs. of  10 variables:
##  $ org   : Factor w/ 1 level "cbsnyt": 1 1 1 1 1 1 1 1 1 1 ...
##  $ survey: int  9158 9158 9158 9158 9158 9158 9158 9158 9158 9158 ...
##  $ bush  : int  NA 1 0 0 1 1 1 1 0 0 ...
##  $ state : int  7 39 31 7 33 33 39 20 33 40 ...
##  $ edu   : int  3 4 2 3 2 4 2 2 4 1 ...
##  $ age   : int  1 2 4 1 2 4 2 4 3 3 ...
##  $ female: int  1 1 1 1 1 1 0 1 0 0 ...
##  $ black : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ region: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ v_prev: num  0.567 0.527 0.564 0.567 0.524 ...
```

```r
head(polls_subset)
```

```
##      org survey bush state edu age female black region    v_prev
## 1 cbsnyt   9158   NA     7   3   1      1     0      1 0.5666333
## 2 cbsnyt   9158    1    39   4   2      1     0      1 0.5265667
## 3 cbsnyt   9158    0    31   2   4      1     0      1 0.5641667
## 4 cbsnyt   9158    0     7   3   1      1     0      1 0.5666333
## 5 cbsnyt   9158    1    33   2   2      1     0      1 0.5243666
## 6 cbsnyt   9158    1    33   4   4      1     0      1 0.5243666
```

IDS 702

1988 ELECTIONS ANALYSIS

```
summary(polls_subset)
```

```
##      org           survey          bush            state            edu       
##  cbsnyt:2193   Min.   :9158   Min.   :0.0000   Min.   : 1.00   Min.   :1.000  
##                1st Qu.:9158   1st Qu.:0.0000   1st Qu.:14.00   1st Qu.:2.000  
##                Median :9158   Median :1.0000   Median :26.00   Median :2.000  
##                Mean   :9158   Mean   :0.5578   Mean   :26.11   Mean   :2.653  
##                3rd Qu.:9158   3rd Qu.:1.0000   3rd Qu.:39.00   3rd Qu.:4.000  
##                Max.   :9158   Max.   :1.0000   Max.   :51.00   Max.   :4.000  
##                               NA's   :178                                     
##       age            female           black             region     
##  Min.   :1.000   Min.   :0.0000   Min.   :0.00000   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:2.000  
##  Median :2.000   Median :1.0000   Median :0.00000   Median :2.000  
##  Mean   :2.289   Mean   :0.5887   Mean   :0.07615   Mean   :2.431  
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000  
##  Max.   :4.000   Max.   :1.0000   Max.   :1.00000   Max.   :5.000  
##                                                                    
##      v_prev      
##  Min.   :0.1530  
##  1st Qu.:0.5278  
##  Median :0.5481  
##  Mean   :0.5550  
##  3rd Qu.:0.5830  
##  Max.   :0.6927  
##                  
```

1988 ELECTIONS ANALYSIS

```
polls_subset$v_prev <- polls_subset$v_prev*100 #rescale
polls_subset$region_label <- factor(polls_subset$region,levels=1:5,
                                    labels=c("NE","S","N","W","DC"))
#we consider DC as a separate region due to its distinctive voting patterns
polls_subset$edu_label <- factor(polls_subset$edu,levels=1:4,
                                 labels=c("No HS","HS","Some College","College Grad"))
polls_subset$age_label <- factor(polls_subset$age,levels=1:4,
                                 labels=c("18-29","30-44","45-64","65+"))
#the data includes states but without the names, which we will need,
#so let's grab that from R datasets
data(state)
#"state" is an R data file (type ?state from the R command window for info)
state.abb #does not include DC, so we will create ours
```

```
##  [1] "AL" "AK" "AZ" "AR" "CA" "CO" "CT" "DE" "FL" "GA" "HI" "ID" "IL" "IN" "IA"
## [16] "KS" "KY" "LA" "ME" "MD" "MA" "MI" "MN" "MS" "MO" "MT" "NE" "NV" "NH" "NJ"
## [31] "NM" "NY" "NC" "ND" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VT"
## [46] "VA" "WA" "WV" "WI" "WY"
```

```
#In the polls data, DC is the 9th "state" in alphabetical order
state_abbr <- c (state.abb[1:8], "DC", state.abb[9:50])
polls_subset$state_label <- factor(polls_subset$state,levels=1:51,labels=state_abbr)
rm(list = ls(pattern = "state")) #remove unnecessary values in the environment
```

IDS 702

1988 ELECTIONS ANALYSIS

```
###### View properties of the data
head(polls_subset)
```

```
##      org survey bush state edu age female black region   v_prev region_label
## 1 cbsnyt   9158   NA     7   3   1      1     0      1 56.66333           NE
## 2 cbsnyt   9158    1    39   4   2      1     0      1 52.65667           NE
## 3 cbsnyt   9158    0    31   2   4      1     0      1 56.41667           NE
## 4 cbsnyt   9158    0     7   3   1      1     0      1 56.66333           NE
## 5 cbsnyt   9158    1    33   2   2      1     0      1 52.43666           NE
## 6 cbsnyt   9158    1    33   4   4      1     0      1 52.43666           NE
##      edu_label age_label state_label
## 1 Some College     18-29          CT
## 2 College Grad     30-44          PA
## 3           HS       65+          NJ
## 4 Some College     18-29          CT
## 5           HS     30-44          NY
## 6 College Grad       65+          NY
```

```
dim(polls_subset)
```

```
## [1] 2193   14
```

IDS 702

1988 ELECTIONS ANALYSIS

```r
###### View properties of the data
str(polls_subset)
```

```
## 'data.frame':    2193 obs. of  14 variables:
##  $ org         : Factor w/ 1 level "cbsnyt": 1 1 1 1 1 1 1 1 1 1 ...
##  $ survey      : int  9158 9158 9158 9158 9158 9158 9158 9158 9158 9158 ...
##  $ bush        : int  NA 1 0 0 1 1 1 1 0 0 ...
##  $ state       : int  7 39 31 7 33 33 39 20 33 40 ...
##  $ edu         : int  3 4 2 3 2 4 2 2 4 1 ...
##  $ age         : int  1 2 4 1 2 4 2 4 3 3 ...
##  $ female      : int  1 1 1 1 1 1 0 1 0 0 ...
##  $ black       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ region      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ v_prev      : num  56.7 52.7 56.4 56.7 52.4 ...
##  $ region_label: Factor w/ 5 levels "NE","S","N","W",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ edu_label   : Factor w/ 4 levels "No HS","HS","Some College",..: 3 4 2 3 2 4 2 2 4 1 ...
##  $ age_label   : Factor w/ 4 levels "18-29","30-44",..: 1 2 4 1 2 4 2 4 3 3 ...
##  $ state_label : Factor w/ 51 levels "AL","AK","AZ",..: 7 39 31 7 33 33 39 20 33 40 ...
```

IDS 702

What's next?

Move on to the readings for the next module!