# IDS 702: Module 4.6

## Multilevel/hierarchical logistic models (illustration)

Dr. Olanrewaju Michael Akande

# 1988 ELECTIONS ANALYSIS RECAP

2193 observations from one of eight CBS News surveys.

| 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) |

The data is in the file `polls_subset.txt` on Sakai.

# 1988 ELECTIONS ANALYSIS RECAP

```
polls_subset <- read.table("data/polls_subset.txt",header=TRUE)
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"))
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+"))
data(state)
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"))
```

# 1988 ELECTIONS ANALYSIS

- I will not do any substantial EDA here.
- I expect you to be able to do this yourself.
- Let's just take a look at the amount of data we have for "bush" and the age:edu interaction.

```
###### Exploratory data analysis
table(polls_subset$bush) #well split by the two values
```

```
##
##    0    1
##  891 1124
```

```
table(polls_subset$edu,polls_subset$age)
```

```
##
##       1   2   3   4
##   1  44  42  67  96
##   2 232 283 223 116
##   3 141 205  99  54
##   4 119 285 125  62
```

# 1988 ELECTIONS ANALYSIS

- As a start, we will consider a simple model with fixed effects of race and sex and a random effect for state (50 states + the District of Columbia).

$$\begin{aligned} \text{bush}_i|\boldsymbol{x}_i &\sim \text{Bernoulli}(\pi_i); \quad i = 1, \ldots, n; \quad j = 1, \ldots, J = 51; \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \gamma_{0j[i]} + \beta_1 \text{female}_i + \beta_2 \text{black}_i; \\ \gamma_{0j} &\sim N(0, \sigma^2_{\text{state}}). \end{aligned}$$

- We can also write

$$\begin{aligned} \text{bush}_i|\boldsymbol{x}_i &\sim \text{Bernoulli}(\pi_i); \quad i = 1, \ldots, n; \quad j = 1, \ldots, J = 51; \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \gamma^{\text{state}}_{0j[i]} + \beta_{\text{female}} \text{female}_i + \beta_{\text{black}} \text{black}_i; \\ \gamma_{0j} &\sim N(0, \sigma^2_{\text{state}}). \end{aligned}$$

- In R, we have

```
library(lme4)
model1 <- glmer(bush ~ black+female+(1|state_label),family=binomial(link="logit"),
data=polls_subset)
summary(model1)
```

# 1988 ELECTIONS ANALYSIS

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: bush ~ black + female + (1 | state_label)
##    Data: polls_subset
##
##      AIC      BIC   logLik deviance df.resid
##   2666.7   2689.1  -1329.3   2658.7     2011
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7276 -1.0871  0.6673  0.8422  2.5271
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  state_label (Intercept) 0.1692   0.4113
## Number of obs: 2015, groups:  state_label, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.44523    0.10139   4.391 1.13e-05 ***
## black       -1.74161    0.20954  -8.312  < 2e-16 ***
## female      -0.09705    0.09511  -1.020    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) black
## black  -0.119
## female -0.551 -0.005
```

# 1988 ELECTIONS ANALYSIS

- Looks like we dropped some NAs.

```
c(sum(complete.cases(polls_subset)),sum(!complete.cases(polls_subset)))
```

```
## [1] 2015  178
```

- Not ideal; we'll learn about methods for dealing with missing data soon.
- Interpretation of results:
  - For a fixed state (or across all states), a non-black male respondent has odds of $e^{0.45} = 1.57$ of supporting Bush.
  - For a fixed state and sex, a black respondent as $e^{-1.74} = 0.18$ times (an 82% decrease) the odds of supporting Bush as a non-black respondent; you are much less likely to support Bush if your race is black compared to being non-black.
  - For a given state and race, a female respondent has $e^{-0.10} = 0.91$ (a 9% decrease) times the odds of supporting Bush as a male respondent. However, this effect is not actually statistically significant!

# 1988 ELECTIONS ANALYSIS

- The state-level standard deviation is estimated at 0.41, so that the states do vary some, but not so much.
- We no longer have a term for residual standard deviation (residual standard error). Why is that?
- I expect that you will be able to interpret the corresponding confidence intervals.

```
## Computing profile confidence intervals ...
```

```
##                   2.5 %      97.5 %
## .sig01        0.2608567  0.60403426
## (Intercept)   0.2452466  0.64871233
## black        -2.1666001 -1.34322366
## female       -0.2837099  0.08919986
```

# 1988 ELECTIONS ANALYSIS

- Let's fit a more sophisticated model that includes other relevant survey factors, such as
  - region
  - prior vote history (note that this is a state-level predictor),
  - age, education, and the interaction between them.
- In R, we have

```
model2 <- glmer(bush ~ black + female + v_prev + edu_label:age_label +
            (1|state_label) + (1|region_label),
            family=binomial(link="logit"),data=polls_subset)
```

```
## fixed-effect model matrix is rank deficient so dropping 1 column / coefficient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0122335 (tol = 0.001, component 1)
```

- Why do we have a rank deficient model?

# 1988 ELECTIONS ANALYSIS

- Also, it looks like we have a convergence issue. This can happen when dealing with multilevel models. We have so many parameters to estimate from the interaction terms `edu_label:age_label` (16 actually), and it looks like that's causing a problem.
- Could be that we have too many $\text{bush}_i = 1$ or $0$ values for certain combinations. You should check!
- Let's treat those as varying effects instead. That is,

$$\begin{aligned}\text{logit}\left(\Pr[\text{bush}_i = 1]\right) &= \beta_0 + \gamma^{\text{region}}_{0m[i]} + \gamma^{\text{state}}_{0j[i]} + \gamma^{\text{age.edu}}_{0k[i],l[i]} \\ &\quad + \beta_{\text{f}}\text{female}_i + \beta_{\text{b}}\text{black}_i + \beta_{\text{v\_prev}}\text{v\_prev}_{j[i]}; \\ \gamma_{0m} \sim N(0, \sigma^2_{\text{region}}), &\quad \gamma_{0j} \sim N(0, \sigma^2_{\text{state}}), \quad \gamma_{0k,l} \sim N(0, \sigma^2_{\text{age.edu}}).\end{aligned}$$

- In `R`, we have

```
model3 <- glmer(bush ~ black + female + v_prev + (1|state_label)
            + (1|region_label) + (1|edu_label:age_label),
            family=binomial(link="logit"),data=polls_subset)
```

- This seems to run fine; we are able to borrow information which helps.

# 1988 ELECTIONS ANALYSIS

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## bush ~ black + female + v_prev + (1 | state_label) + (1 | region_label) +
##     (1 | edu_label:age_label)
##    Data: polls_subset
##
##      AIC      BIC   logLik deviance df.resid
##   2644.0   2683.3  -1315.0   2630.0     2008
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8404 -1.0430  0.6478  0.8405  2.7528
##
## Random effects:
##  Groups              Name        Variance Std.Dev.
##  state_label         (Intercept) 0.03768  0.1941
##  edu_label:age_label (Intercept) 0.02993  0.1730
##  region_label        (Intercept) 0.02792  0.1671
## Number of obs: 2015, groups:
## state_label, 49; edu_label:age_label, 16; region_label, 5
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.50658    1.03365  -3.392 0.000693 ***
## black       -1.74530    0.21090  -8.275  < 2e-16 ***
## female      -0.09956    0.09558  -1.042 0.297575
## v_prev       0.07076    0.01853   3.820 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr) black  female
## black  -0.036
## female -0.049 -0.004
## v_prev -0.992  0.027 -0.006
```

# 1988 ELECTIONS ANALYSIS

- Remember that in the first model, the state-level standard deviation was estimated as 0.41. Looks like we are now able to separate that (for the most part) into state and region effects.
- Interpretation of results:
  - For a fixed state, education and age bracket, a non-black male respondent with zero prior average Republican vote share, has odds of $e^{-3.51} = 0.03$ of supporting Bush (no one really has 0 value for `v_prev`).
  - For a fixed state, sex, education level, age bracket and zero prior average Republican vote share, a black respondent has $e^{-1.75} = 0.17$ times (an 83% decrease) the odds of supporting Bush as a non-black respondent, which is about the same as before.
  - For each percentage point increase in prior average Republican vote share, residents of a given state, race, sex, education level age bracket have $e^{0.07} = 1.07$ times the odds of supporting Bush.

# 1988 ELECTIONS ANALYSIS

- Due to the number of categories, the inference in the frequentist model is not entirely reliable as
  - it does not fully account for uncertainty in the estimated variance parameters, and
  - it uses an approximation for inference.
- We can fit the model under the Bayesian paradigm in the `brms` package, using mildly informative priors and quantify uncertainty based on the posterior samples.
- Windows users: install Rtools for windows, then the `rstan` package in R.
- Mac users: install Xcode, open it to accept the license agreement, then open R/RStudio and install the `rstan` package.
- In-class analysis: move to the R script here.

# What's next?

Move on to the readings for the next module!