BIOS 755: Multilevel Logistic Regression

Alexander McLain

April 5, 2023

Introduction

- Multilevel modeling can be applied to logistic regression and other generalized linear models
- This will be similar to the linear case where coefficients and random effects will be grouped at different levels in the data.
- The only difference is how level 1 error, i.e., the residual, is captured.

Examples of GLMM

- Binary logistic model with random intercepts:

$$\mathrm{logit}\{P(Y_{ij}=1|b_j)\} = \beta_0 + \beta_1 X_i + b_j$$

$$\log\left\{\frac{P(Y_{ij}=1|b_i)}{1-P(Y_{ij}=1|b_j)}\right\} = \beta_0 + \beta_1 X_j + b_j$$

$$P(Y_{ij}=1|b_i) = \frac{e^{\beta_0+\beta_1 X_j + b_j}}{1+e^{\beta_0+\beta_1 X_j + b_j}}$$

  with $b_i \sim N(0, \sigma^2)$.

- Here, e^{β_1} gives the level 1 change in the odds for a 1-unit increase in X.

Examples of GLMM

- Random coefficients Poisson regression model:

$$\log\{E(Y_{ij}|\boldsymbol{b}_j)\} = \log(n_{ij}) + \beta_0 + \beta_1 G_i + \beta_2 X_{ij} + \beta_3 G_i X_{ij} + b_{j0} + b_{j1} X_{ij}$$

$$\log\left\{E\left(\frac{Y_{ij}}{n_{ij}}\middle|\boldsymbol{b}_j\right)\right\} = \beta_0 + \beta_1 G_i + \beta_2 X_{ij} + \beta_3 G_i X_{ij} + b_{j0} + b_{j1} X_{ij}$$

$$E\left(\frac{Y_{ij}}{n_{ij}}\middle|\boldsymbol{b}_j\right) = \exp(\beta_0 + \beta_1 G_i + \beta_2 X_{ij} + \beta_3 G_i X_{ij} + b_{j0} + b_{j1} X_{ij})$$

  where $\boldsymbol{b}_j = (b_{j0}\ b_{j1})$ with a random intercept and a random effect of X and $\boldsymbol{b}_j \sim N(0, \boldsymbol{G})$.
- Here, e^{β_1} gives the level 1 change in the rate (i.e., $E(Y_{ij})$ per each n_{ij}) between treatment and control groups.

Example

Guatemalan immunization campaign

- Data are available from the National Survey of Maternal and Child Health conducted in Guatemala in 1987
- A nationally representative sample of 5160 women aged between 15 and 44 were interviewed
- The questionnaire included questions determining the immunization status of children who were born in the previous 5 years and alive at the time of the interview

Example

- Beginning 1986, the Guatemalan government undertook a series of campaign to immunize the population against major childhood diseases
- An important explanatory variable is whether the child was at least 2 years old at the time of the interview, in which case the child was old enough to be immunized during the 1986 campaign.
- If this variable is associated with immunization, there is some indication that the government campaign worked.

What are the levels of data?

Two-level model

- As we discussed last week, a two level model would be similar to what we've done before
$$logit\{P(Y_{ij} = 1|\boldsymbol{X})\} = \beta_0 + \beta_1 X_{ij} + b_{j0}$$
where
  - Y_{ij} is the immunization status for the ith child from the jth mother
  - X_{ij} is the indicator that the child is at least 2 years old.
  - $b_{j0} \sim N(0, \sigma_b)$ is the mother level random intercept
- **Question:** how do we estimate the ICC?

Estimating the ICC (two-level)

- To estimate the ICC we need to put the logistic model into a latent variable formulation.
- In this model we assume underlying the observed dichotomous response (whether the child was immunized), there is an unobserved or latent continuous response.
- This latent response represents the propensity to be immunized.
- If this latent response is greater than zero, then the observed response is 1 else the response is 0.

Estimating the ICC (two-level)

- The latent variable formulation is

$$Y_{ij}^* = \beta_0 + \beta_1 X_{ij} + b_{j0} + \varepsilon_{ij}$$

where

$$\begin{aligned} Y_{ij}^* > 0 \quad &\rightarrow \quad Y_{ij} = 1 \\ Y_{ij}^* \leq 0 \quad &\rightarrow \quad Y_{ij} = 0 \end{aligned}$$

and $E(\varepsilon_{ij}) = 0$

Estimating the ICC (two-level)

- In logistic regression the error ε_{ij} is assumed to have a logistic distribution where

$$\Pr\left(\varepsilon_{ij} < \tau | \boldsymbol{X}_{ij}\right) = \frac{\exp(\tau)}{1+\exp(\tau)}$$

  and $\mathrm{Var}\left[\varepsilon_i | x_i\right] = \frac{\pi^2}{3} \approx 3.29$
- The estimate the ICC for a two-level model is

$$\mathrm{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi^2}{3}} \approx \frac{\sigma_b^2}{\sigma_b^2 + 3.29}$$

Median Odds Ratio

- Similar to the ICC, the Median Odds Ratio (MOR) is a quantification of clustering for logistic regression.
- The goal of the MOR is to give the amount of clustering using the scale of ORs.
- Suppose that two kids have equal predictors variables but are from different mothers. The OR of immunization between these kids is:

$$OR_{jk} = \frac{P(Y_{ij} = 1)}{P(Y_{i'k} = 1)} = e^{b_j - b_k} \tag{1}$$

  since the random effects are the only difference.
- The idea of MOR is to calculate all the possible ORs obtained from (1) for all j and k such that $b_j > b_k$.

Median Odds Ratio

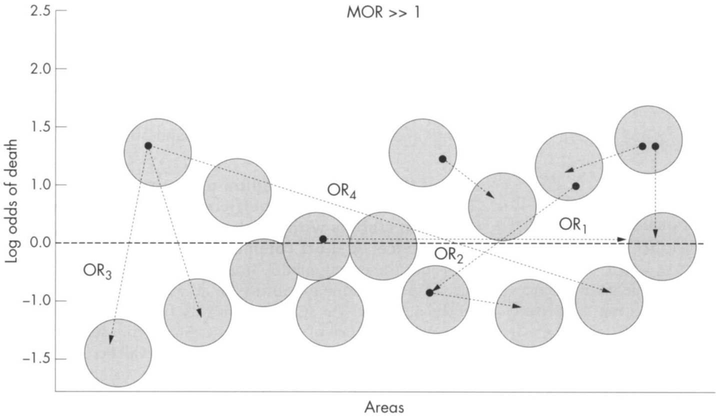

Figure from: Merlo, J., et al. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health*, 60(4), 290-297.

Median Odds Ratio

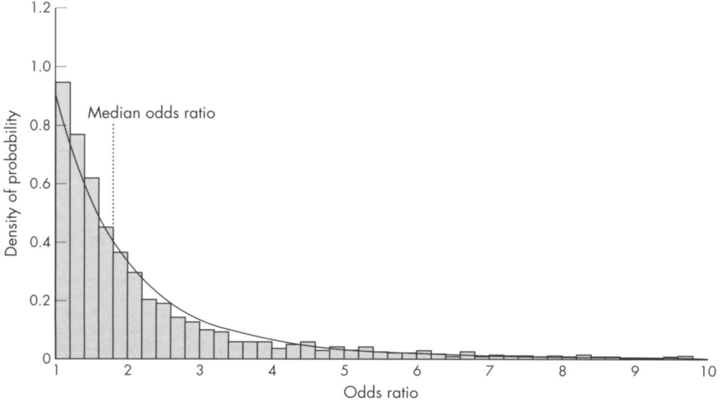

Figure from: Merlo, J., et al. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health*, 60(4), 290-297.

Median Odds Ratio

- Calculating the MOR is straightforward:

$$MOR = \exp\left(0.954\sqrt{\sigma_b^2}\right) = \exp(0.954\sigma_b)$$

- If MOR $= 1$, then there would be no differences between mothers in the probability of being immunized.
- If $MOR = 1.8$ then the median difference between mothers increased the child level odds of being immunized by 80% when randomly picking out two mothers.
- That is, if a child was to randomly change to a different mother that had higher immunization probability, the median increase in the child's odds of immunization is by a factor of 1.8.
  - So, 50% of the time it would be lower than 1.8, 50% of the time it would be higher than 1.8.

Three-level model

- In our example, there is actually a third level (community).
- It is of interest to fit the model

$$logit\{P(Y_{ijk} = 1|\boldsymbol{X})\} = \beta_0 + \beta_1 X_{ijk} + b_{jk0} + b_{k0}$$

where

  - Y_{ijk} is the immunization status for the ith child from the jth mother in the kth community
  - X_{ijk} is the indicator that the child is at least 2 years old
  - $b_{jk0} \sim N(0, \sigma^2_{(2)})$ is the mother-level random intercept
  - $b_{k0} \sim N(0, \sigma^2_{(3)})$ is the community-level random intercept

Estimating the ICC (three-level)

- Correlation across mothers within the same community

$$\rho(comm) = corr(Y^*_{ijk}, Y^*_{i'j'k}) = \frac{\sigma_3^2}{\sigma_2^2 + \sigma_3^2 + \pi^2/3}$$

- Correlation across children for the same mother (ignoring community) is

$$\rho(mother) = corr(Y^*_{ijk}, Y^*_{i'jk}) = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_3^2 + \pi^2/3}$$

- Correlation across children for the same mother and within the same community

$$\rho(mother, comm) = corr(Y^*_{ijk}, Y^*_{i'jk}) = \frac{\sigma_2^2 + \sigma_3^2}{\sigma_2^2 + \sigma_3^2 + \pi^2/3}$$

Estimating the MOR (three-level)

- MOR for changing mothers and staying in the same community

$$MOR(comm) = \exp(0.954\sigma_3)$$

- MOR for changing communities and mother staying with the same

$$MOR(mother) = \exp(0.954\sigma_2)$$