Assignment 2
Longitudinal Data (PH242C: STAT247C)

1 Notation

Consider a clustered, longitudinal study of diarrheal disease in children in Ecuador [Eisenberg et al., 2006]. Though not precisely how this study was actually done, for this assignment, assume the study used a) a random sample of 10 *independent* villages, b) within those villages, a random sample of 35 households that contained children under the age of 5, and c) within the households, all children under the age of 5. At three annual visits, households data were collected on:

1. the (binary) outcome of diarrhea over the last week for each child,
2. whether or not any water treatment was currently being used at the household level, and
3. the proportion of villagers that traveled to the closest main town over the past month (village-level "movement" variable).

Keeping as close as you can to the notation in Chapter 1, tell us

- What is the value of m (number of independent units)?
- What is a good notation for the outcome, Y, including the appropriate indices, and the range of each index, e.g., $i = 1, ..., m, j = 1, ...,$? (This may take more subscripts that we've used in examples in class.)
- Provide a logistic model that relates the probability of diarrhea for a particular child at one of the visits given the two explanatory variables listed above. Assume that all observations have the same relationship (coefficients) with the explanatory variables.

- Expand this model to allow the associations of water treatment to differ randomly by household, and the association of movement to vary randomly by village.

  (However you do this, be explicit about the definitions of your notation)

2 Data-generating distributions for repeated measures data

- Question 1.4 in chapter 1 (Jewell and Hubbard).
- Simulate data in STATA from the model implied by the data-generating description in Question 1.4. Assume the random variables are normally distributed (every thing else is provided). Also, assume the following parameters:
  - $\sigma_\alpha^2 = 0.5$
  - $cor(Y_{ij}, Y_{ij'}) \equiv \rho = 0.3$
  - $\mu = EY_{ij} = 10$
  - Simulate data at sample sizes of $m = 20, 100$ and 1000 always with $n_1 = n_2 = \ldots n_m = 2$. For each of these simulations, estimate ρ and μ. Turn in the following:

  1. Stata code used to generate simulation and estimates of ρ and μ.
  2. Plot of these estimates versus the sample size, m, separately for ρ and μ including putting a horizontal line for the true value of these.
  3. Short explanation of what these plots show.

3 References

Joseph NS Eisenberg, William Cevallos, Karina Ponce, Karen Levy, Sarah J Bates, James C Scott, Alan Hubbard, Nadia Vieira, Pablo Endara, Mauricio Espinel, et al. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural ecuador. Proceedings of the National Academy of Sciences, 103(51):1946019465, 2006.