1. **Simulating null distributions.** Consider the following HLM, model M_1:

$$\begin{aligned} y_{i,j} &= \beta_0 + \beta_1 x_{i,j} + \beta_2 w_j + b_{0,j} + b_{1,j} x_{i,j} + \epsilon_{i,j} \\ \{\mathbf{b}_j\} &\sim \text{ iid } N(\mathbf{0}, \Psi) \\ \{\epsilon_{i,j}\} &\sim \text{ iid } N(0, \sigma^2) \end{aligned}$$

where

$$\text{Cov}[\mathbf{b}_j] = \Psi = \begin{pmatrix} \psi_0^2 & \psi_{01} \\ \psi_{01} & \psi_1^2 \end{pmatrix}.$$

(a) Identify in the model

- macro explanatory variables and micro explanatory variables;
- fixed effect parameters;
- random effects;
- variance and covariance parameters.

Explain in words why we wouldn't include a term like $b_{2,j} w_j$ in the model.

(b) Let M_0 refer to the model in which $\psi_1^2 = 0$ (and so $\psi_{01} = 0$ as well). What is the (asymptotic) null distribution of the LRT statistic for testing M_0 versus M_1? How would you obtain a p-value based on this null distribution?

(c) Do a simulation study, simulating datasets under the above model with $m = 20$, $n = 20$, $\beta_0 = \beta_1 = \beta_2 = \sigma^2 = \psi_0^2 = 1$, $\{x_{i,j}\}$ and $\{w_j\}$ both being i.i.d. standard normal, and $\psi_1^2 = \psi_{01} = \psi_{10} = 0$. In other words, you are simulating data under a null model M_0 in which there is no across group heterogeneity in slopes with $x_{i,j}$. Simulate several thousands of such datasets. For each simulated dataset,

i. Fit the full model M_1 and the null model M_0 via maximum likelihood, and obtain the likelihood ratio test statistic.

ii. Obtain the p-value for testing M_0 versus M_1.

You should now have thousands of LRT statistics and p-values. Make a histogram of the LRT statistics and compare it to the null distribution you identified in (b). Make a histogram of the p-values and describe the distribution.

(d) Repeat (c) using different values of the parameters, and describe your results.

2. **Earthquakes.** Read in the `Earthquake` data from homework directory using the `dget` command. For any given earthquake, vertical acceleration data (`accel`) was gathered at potentially several locations, with the distance of each location from the epicenter of the quake also being recorded. Also, the soil type was recorded for each location, as was the magnitude of each quake on the Richter scale.

   (a) Fit an HLM for `accel` as a function of the other variables, with fixed and random effects for all micro-level covariates, and fixed effects only for all macro-level covariates. Describe the results of the model fitting routine, and evaluate normality of the residuals using a qqplot.

   (b) Plot `accel` versus `distance`, pooled across groups. Does the relationship look linear like the model assumes? Obtain a transformation of both variables so that the relationship looks approximately linear.

   (c) Refit the HLM from (a) but with `accel` and `distance` replaced by your transformed variables. Evaluate the residuals with a qqnorm plot and compare to the results from (a).

   (d) Evaluate using two LRTs whether or not there is across-quake heterogeneity in the slopes with `soil` and `tdist`, where `tdist` is your transformed distance variable. Specifically identify your null distributions and your p-values. Also evaluate the fixed effects using LRTs.

   (e) Write a couple of paragraphs describing your findings, in terms of the relationship between `accel` and the explanatory variables. Make sure you describe the directionality of the fixed effects coefficients.