Exercise Day 1

Load the river-blindness (RB) data `LiberiaRemoData.csv` into R.

1. Fit a Binomial for the nodule prevalence p_i such that
$$\log\left\{\frac{p_i}{1-p_i}\right\} = \beta_0 + \beta_1 x_{1,i} + \beta x_{2,i}$$
where $(x_{i,1}, x_{i,2})$ are the UTM coordinates of the i-th sampled village. Based on the estimates of β_1 and β_2, in which part of Liberia higher values of prevalence are found?

2. Based on the model estimated in the previous point, predict the nodule prevalence using a 3 by 3 km regular grid covering the whole of Liberia.

3. Generate 1,000 samples for the estimates of β_1 and β_2 using a Gaussian approximation to distribution the maximum likleihood estimator. Use these to compute 1) the standard errors of the estimated prevalence and 2) the probability that nodule prevalence exceeds 20%. Display the resulting map as a raster file using the `tmap` and `raster` packages. Which areas are at least 80% likely to exceed 20% prevalence?

4. Repeat 1, 2 and 3, using elevation as a spatial predictor, i.e.
$$\log\left\{\frac{p_i}{1-p_i}\right\} = \beta_0 + \beta_1 \log d(x_i)$$
where $d(x_i)$ is the elevation in meters a location x_i.

5. Which of the models in 1 and 3 best fit the data?

6. **Challenge question.** Use the best model you selected to generate estimates at district level (ADM 2) and provide 95% confidence intervals. Finally, generate a map using the `tmap` package.

7. Fit a generalized linear mixed to the prevalence data, using elevation to give a linea predictor
$$\log\left\{\frac{p_i}{1-p_i}\right\} = \beta_0 + \beta_1 \log d(x_i) + Z_i$$
where Z_i are i.i.d. Gaussian variable with mean 0 and variance σ^2. What is the estimate for σ^2 and what does it mean?

8. Perform the test for residual spatial correlation based on the variogram using the model from the previous question. What do you conclude?