Common Analysis Problems

August 19, 2015

1. You are interested in modeling the relationship between per capita income and an index of air pollution (i.e., the Kuznets effect). You hypothesize that air pollution increases then declines as per capita income increases. You have data on the mean (and the standard deviation for the mean) for each county's air quality index (a continuous non-negative variable). The predictor variable (income) is not measured perfectly because it is based on a sample – you have a mean and its standard deviation for the per capita income of each county. How would you analyze these data?

2. You have data on the relationship between incidence of lung cancer in households (1 if cancer is present and 0 if no cancer) and radon levels in the house for 1000 houses within 40 counties within a state. You also have data on the clay soil content at the county level, which you heroically assume is know without error. How would you model the effect of radon and soil type on the probability on the probability of lung cancer?

3. You have plot level data on diversity of plant communities. The data consist of counts y_{ij} of the number of individuals of species i on $j = 1...J$ same-sized plots, and the total number of individuals on plot j is reported as n_j. How would you estimate an index (H) of species diversity across the community, where $H = -\sum_{i=1}^{R} p_i \ln\big(p_i\big)$, (p_i) is the mean proportion of the i_{th} species in in the community, and R is the total number of species present? All estimates must include proper accounting for uncertainty.

4. You have data on the number of willow seedings that establish on 100 different 10 × 10 meter plots. Assume these data are measured perfectly, i.e. you did not over or under count seedlings. You also have 5 measurements of soil water and 1 measurement of percent cover (estimated visually) on each plot. How would you model the effect of soil water and percent bare soil on the number of plants established? You also have paired data of visually estimated cover and actual cover on 15 calibration plots (separate from the 100 plots with seedling and soil water data), where the actual proportion of vegetated soil was laboriously measured using 1 × 1 meter gridded quadrats. How would you incorporate uncertainty in the percent cover and soil water covariates in your analysis?

5. Now presume that the 100 plots are arranged in 5 different stream reaches. You have data on peak runoff in each of the reaches, which you may assume is measured perfectly. You want to understand variation at the catchment scale created by peak runoff. How would you include these data in the model you developed in problem 4?