Assignment 1

Sam Olson

Preface

For each of the following situations, define random variables that might be appropriate for conducting an analysis. If distributions are to be assigned to these variables, what should be the support of those distributions? Would it be reasonable to consider the variables to be independent? In some cases there may not be a single correct answer. While it is certainly reasonable to be thinking about some type of model you are not being asked to identify a model structure that you would use to address the objectives identified. You are only being asked to define appropriate random variables (and perhaps covariates), identify what support a distribution assigned to them should have, and make a preliminary assessment about an assumption of independence. Do not suggest a model, you will loose points if you do.

Q1

1. (20 pt.) As part of its Long Term Resource Monitoring Program, the USGS Upper Midwest Environmental Sciences Center located in LaCrosse, Wisconsin has sampled sediment from the Upper Mississippi River from about 120 sites at each of six reaches (or stretches) of the river, from 1992 to 2004. In each stretch of the river, samples were taken from a number of primary habitat categories used as strata in a sampling design. Those categories were Backwater, Impounded Water, Side Channel, and Main Channel Border. Sampling locations were selected separately each year so that repeated sampling of the same location over time did not occur. The sediment samples are brought back to the laboratory, run through a sieve, and the types and numbers of invertebrates are recorded, as is the specific predominant sediment type of sand, silt,or clay (there are actually 6 categories, but 3 is enough to get the idea). Water depth is also measured at the time each sample is collected.

The USGS would like to use these data to address a number of questions related to the status of mayflies (Ephemeroptera) in the river. Mayflies form the basis for a number of aquatic food chains and are also generally considered an indicator of water quality (rather, their absence is considered an indicator of a lack of water quality). There are any number of objectives that might be identified for this study. Here, we will be concerned with only two. Restrict your answer to issues that are relevant for these two objectives.

(a)

Is the presence/absence of mayflies at sampling locations related to the primary habitat category and/or the specific sediment type?

(b)

Has the abundance of mayflies exhibited a systematic change over the period 1992 to 2004?

Answer

Define appropriate random variables (and perhaps covariates), and what support a distribution assigned to them should have:

A R.V. for the presence of mayflies (in a sample),
$Y_1 : \Omega \to \{0, 1\}, \quad Y_1(\omega) \in \{0, 1\}$

A R.V. for the number of mayflies (in a sample),
$Y_2 : \Omega \to \mathbb{N}_0, \quad Y_2(\omega) \in \mathbb{N}_0$

Since presence implies at least one mayfly, we may also write the joint support as:

$$(Y_1, Y_2) \in \{(0, 0)\} \cup \{(1, y) : y \in \mathbb{N}_0^+\}, \quad \text{that is, } Y_1 = \mathbb{I}\{Y_2 > 0\},$$

where $\mathbb{I}$ is the indicator function.

(Covariate) A R.V. for primary habitat category,
$X_1 : \Omega \to \{\text{Backwater}, \text{Impounded Water}, \text{Side Channel}, \text{Main Channel Border}\}$

(Covariate) A R.V. for sediment type (3 shown for brevity),
$X_2 : \Omega \to \{\text{Sand}, \text{Silt}, \text{Clay}, \dots\}$

(Covariate) A R.V. for year,
$X_3 : \Omega \to \{1992, \dots, 2004\}$

(Covariate) A R.V. for number of invertebrates recorded in the sample (total across types), $X_4 : \Omega \to \mathbb{N}_0$

(Covariate) A R.V. for water depth at the time of sampling, $X_5 : \Omega \to [0, \infty)$

(Covariate) A R.V. for reach (river stretch), $X_6 : \Omega \to \{1, 2, 3, 4, 5, 6\}, \quad X_6(\omega) \in \{1, 2, 3, 4, 5, 6\}$

(Covariate) Site identifier (within reach–year), $S : \Omega \to \{1, 2, \dots, n_{ry}\}, \quad S(\omega) \in \{1, 2, \dots, n_{ry}\}$,
where n_{ry} is the number of sampled sites in a given reach–year.

Preliminary assessment about an assumption of independence:

Independence of mayfly outcomes across samples *within a year* is unlikely due to spatial correlation (nearby sites may share environmental/morphological conditions, leading to clustering). *Between years*, independence is more plausible because sites are reselected annually, with no sites replicated year-over-year, though the prior point still stands. Also, regarding *independence between random variables*, Y_1 and Y_2 are not independent because of the (joint) support constraint $Y_1 = \mathbb{I}\{Y_2 > 0\}$. Independence is more plausible between other combinations of random variables however, though not guaranteed.

Q2

2. (10 pt.) A study was conducted at the College of Veterinary Medicine at ISU to examine the efficacy of a vaccine for Porcine Reproductive and Respiratory Syndrome (PRRS) virus. This virus infects sows but affects piglets. In the study, 12 pregnant adult sows were given the vaccine and another 12 were not. Both groups were "challenged" (i.e., exposed to the virus). Sows were housed separately. The number of piglets born normal, born weak, and born still born were recorded for each sow. The objective of analysis was to determine whether the vaccine was effective in reducing the effects of the PRSS virus.

Answer

Define appropriate random variables (and perhaps covariates), and what support a distribution assigned to them should have:

Some R.V.s for sow i's piglet outcomes (counts):

A R.V. for the number of normal piglets born to sow i, $Y_{1i} : \Omega \to \mathbb{N}_0$

A R.V. for the number of weak piglets born to sow i, $Y_{2i} : \Omega \to \mathbb{N}_0$

A R.V. for the number of stillborn piglets born to sow i, $Y_{3i} : \Omega \to \mathbb{N}_0$

To describe the joint support: Let N_i be the litter size of sow i (total piglets delivered), then the joint support is given by:

$$(Y_{1i}, Y_{2i}, Y_{3i}) \in \{(y_1, y_2, y_3) : y_k \in \mathbb{N}_0,\ y_1 + y_2 + y_3 = N_i\}$$

Where N_i is a R.V. for the number of piglets born to sow i, $N_i : \Omega \to \mathbb{N}_0, \quad N_i(\omega) \in \mathbb{N}_0$.

(Alternatively) We could also have a R.V. for the type of piglet j born to sow i,

$$Y_{ij} : \Omega \to \{\text{Normal},\ \text{Weak},\ \text{Stillborn}\}, \quad j = 1, \ldots, N_i,$$

With support given by the set $\{\text{Normal},\ \text{Weak},\ \text{Stillborn}\}$ and where N_i denotes the number of piglets born to sow i (again).

(Covariate) A R.V. for vaccination status of sow i, $X_{1i} : \Omega \to \{0, 1\}, \quad X_{1i}(\omega) \in \{0, 1\}$, where $1 =$ vaccinated, $0 =$ not vaccinated.

Preliminary assessment about an assumption of independence:

Independence *between sows* seems reasonable since they are housed separately and treatment is applied at the sow level (the sampling unit/experimental unit for this experiment). Independence *within a litter* (across piglets) is generally not reasonable due to shared sow, consequently vaccination status, and other potential factors (though these factors are not noted above as random variables). Additionally, regarding independence *between random variables* Y_{1i}, Y_{2i}, Y_{3i}, these random variables are not independent because they must sum to N_i (a support constraint).

Q3

3. (10 pt.) A problem considered by Dr. Dixon about 15 years ago was the topic of a seminar he presented to the department. This problem involved the capture of insects by a predatory plant species, a member of the family of pitcher plants Sarraceniaceae. These plants have a long central tube with a hood-shaped part at the upper end. The tube has hair-like structures that point downward. Insects that enter the tube are not able to move back up because of these hairs, and eventually are digested by enzymes in the plant. Insects that do not enter the tube may obtain nectar from the plant without becoming plant food. A primary prey species of the pitcher plant species involved in this study are a certain type of small wasp. The study was designed to determine how effective these plants are at capturing wasps.

The study consisted of two parts. The first part involved direct observation of the plants for several hundred hours. The data recorded were the number of wasps visiting the plants and the number of these that were captured by the plants. There were a total of 376 "plant-hours" of observation, 157 visits, and 2 captures. The second part of the study involved cleaning out a number of plants, leaving the study site, and returning about 2 weeks later (it takes the plants longer than 2 weeks to totally digest a wasp that is captured). The data recorded in this part of the study were the number of wasps captured by the plants over a period of 2 weeks, which was equivalent to 1416 "plant-hours". There were a total of 6 wasps captured in this indirect observation portion of the study.

Our concern here is not with the actual numbers that resulted from this study, but rather with defining random variables that might be used in a statistical analysis. The focus of the seminar by Dr. Dixon was how information from the indirect observation part of the study could be combined with information from the direct observation part of the study to improve estimation of the rate of visits by wasps and the probability of capture given a visit, which were the objectives of the study.

Answer

Define appropriate random variables (and perhaps covariates), and what support a distribution assigned to them should have:

A R.V. for the number of wasp visits to plant p during an observed plant–hour, $V : \Omega_{\text{PH}} \to \mathbb{N}_0, \quad V(\omega) \in \mathbb{N}_0$.

A R.V. for the number of captures during an observed plant–hour, $C : \Omega_{\text{PH}} \to \mathbb{N}_0, \quad (C, V)(\omega) \in \{(c, v) : v \in \mathbb{N}_0,\ c \in \{0, 1, \ldots, v\}\}$
(i.e., the maximum number of captures is constrained by the total number of wasp visits, $0 \leq C \leq V$).

(Alternatively) A R.V. for capture during an observed plant–hour (per visit), $Z_j : \Omega_{\text{visit}} \to \{0, 1\}, \quad Z_j(\omega) \in \{0, 1\}, \quad j = 1, \ldots, V(\omega)$.

A R.V. for the number of captures accumulated over an indirectly observed plant–window, $K : \Omega_{\text{PW}} \to \mathbb{N}_0, \quad K(\omega) \in \{0, 1, \ldots, \sum_{h \in \mathcal{H}} V_h\}$,
where $\mathcal{H}$ is the set of plant–hours aggregated by that window and $\sum_{h \in \mathcal{H}} V_h$ is the total number of visits across those hours.

(Covariate) A R.V. for observation mode, $X_1 : \{\Omega_{\text{PH}}, \Omega_{\text{PW}}\} \to \{\text{direct}, \text{indirect}\}, \quad X_1(\omega) \in \{\text{direct}, \text{indirect}\}$.

Unsure if these should be added

(Covariate) A R.V. for plant identifier, $X_2 : \{\Omega_{\text{PH}}, \Omega_{\text{PW}}\} \to \{\text{plant IDs}\}, \quad X_2(\omega) \in \{\text{plant ID 1, plant ID 2}, \ldots\}$ (number of plants not specified).

(Covariate) A R.V. for exposure time (hours observed for that unit), $X_3 : \{\Omega_{\text{PH}}, \Omega_{\text{PW}}\} \to [0, \infty), \quad X_3(\omega) \in [0, \infty)$.

Preliminary assessment about an assumption of independence:

Within an hour, the random variables C and V are not independent because of the support constraint $0 \leq C \leq V$. Across units, independence *between different plants* is plausible if plants are spatially separated (say, by pot/plot). And *within the same plant* across hours, independence is suspect due to temporal clustering of wasp activity and persistent plant/hour conditions. For the indirect window counts K, dependence on the underlying $\{V_h\}$ and number of captures is expected (independence violated) when windows overlap in plant–time with direct observations; when windows and direct observations are disjoint sets, however, independence is more plausible.