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?

Answer

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

To start, let i denote the sampling site, and let j denote the sampling year, where $i = 1, ..., n_j$, $j = 1992, ..., 2004$, and where n_j denotes the n-th sampling site on the j-th year. (I believe we could also change the indexing of j to start at 0 or 1, though this decision is arbitrary.)

We may then define the R.V.s as follows:

Let $Y_{i,j}$ denote the R.V. for the presence of mayflies at sampling site i on year j. The R.V.s are then defined as:

$$Y_{i,j} = \begin{cases} 1 & \text{if mayflies are present at site i and year j} \\ 0 & \text{o.w.} \end{cases}$$

The support of the R.V.s is given by: $\Omega_{Y_{i,j}} = \{0, 1\}$, where:

$$Y_{i,j}(\omega) \in \{0, 1\} \quad \forall i, j$$

(Covariates): For the purposes of this setup, we may consider "primary habitat category" and the "specific sediment type" covariates. Also, since these are observed (realized) quantities, we consider them non-random.

Note: While we can consider "water depth", "stretch", and "number of invertebrates" to be covariates as well, I would not explicitly include them for the purposes of this question because of the note to `"Restrict your answer to issues that are relevant for these two objectives."`.

Preliminary assessment about an assumption of independence:

Independence of mayfly presence 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 re-selected annually, with no sites replicated year-over-year. Additionally, there may be independence between the noted covariates (habitat and sediment types), but I do not have scientific expertise or theory to substantiate this point. It may be the case that particular sediment types may be correlated with specific habitat categories (and vice versa).

(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:

Keeping the same convention from part (a), let $j = 1992, \ldots, 2004$, and let $i = 1, \ldots, n_j$ where n_j is the number of samples taken in year j.

Then, for a single reach, we may then define the R.V.s for the sample-level counts of mayflies in a sample by:

$$X_{i,j} \in \{0, 1, 2, \ldots\}$$

The support of each $X_{i,j}$ then is $\Omega_{X_{i,j}} = \mathbb{N}_0$, where $X_{i,j}(\omega) \in \mathbb{N}_0$.

Using these random variables, we can construct another R.V. for average mayfly count in a sample *within a year*, again specific to a particular reach. The specific construction may be defined by:

$$Y_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j}$$

The support of each of the Y_j's then is $\Omega_{Y_j} = [0, \infty)$, and $Y_j(\omega) \in [0, \infty)$.

We would then extend the above R.V. constructions to include each of the six reaches (for the sake of subscript headaches, this was avoided).

Note: The above uses a continuous support despite being constructed as a linear function of discrete R.V.s $X_{i,j}$'s. Given the order of magnitude of 120 sites per reach, and the spacing between possible values small ($\frac{1}{n_j} \approx \frac{1}{120} = 0.008$), I would argue this treatment (continuous support) is reasonable.

(Covariates): More of an additional note: Time is treated as an index, not as a covariate in the above.

Preliminary assessment about an assumption of independence:

It seems reasonable to assume independence *across years* (between the R.V.s Y_j), though temporal dependence is possible because of repeated measurements (from the same reach).

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):

Let i denote the sows, where $i = 1, ..., 24$ (2 groups of 12 sows), and j denote the piglets born from sow i, where $j = 0, 1, ..., n_i$, and where n_i denotes the total number of piglets born from sow i.

We may then construct R.V.s for whether piglet j born from sow i was born normal as follows:

$$N_{i,j} = \begin{cases} 1 & \text{if piglet j from sow i is born normal} \\ 0 & \text{otherwise} \end{cases}$$

Similarly we may construct R.V.s for whether piglet j born from sow i was born weak as follows:

$$W_{i,j} = \begin{cases} 1 & \text{if piglet j from sow i is born weak} \\ 0 & \text{otherwise} \end{cases}$$

And also, we may construct R.V.s for whether piglet j born from sow i was stillborn as:

$$S_{i,j} = \begin{cases} 1 & \text{if piglet j from sow i is stillborn} \\ 0 & \text{otherwise} \end{cases}$$

The support of each of the R.V.s is given by: $\Omega = \{0, 1\}$, where:

$$N_{i,j}(\omega) \in \{0,1\} \quad W_{i,j}(\omega) \in \{0,1\} \quad S_{i,j}(\omega) \in \{0,1\} \quad , \forall i, j$$

As defined, for each piglet–sow pair (i, j), exactly one of $N_{i,j}, W_{i,j}, S_{i,j}$ equals 1; that is, the three R.V.s are mutually exclusive and collectively exhaustive, with $N_{i,j} + W_{i,j} + S_{i,j} = 1$, $\forall i, j$.

Note: We also may construct random variable(s) not as counts, but as proportions, e.g. R.V.s for the proportions of piglets born to sow i that were normal, weak, or stillborn. What's more, we could also instead construct R.V.s to "indicate" the status of a piglet j born to sow i.

(Covariate) Vaccination status of sow (i) may be treated as a covariate; again, this is not explicitly random in the context of the experiment, but known/realized.

Preliminary assessment about an assumption of independence:

Independence *between sows* seems reasonable since they are housed separately and treatment (the vaccine) is applied at the sow level (the sampling unit/experimental unit for this experiment). Independence *within a litter* (across piglets for a given sow) is not reasonable though because the piglets share a sow, and consequently share housing condition and vaccination status. Additionally, for each piglet-sow combination (i, j), the three R.V.s being mutually exclusive and collectively exhaustive means they cannot be independent.

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 what support a distribution assigned to them should have):

Given the setup, the R.V.s will be divided into direct and indirect observation.

Respectively then, let us construct a R.V. for number of wasp visits (direct observation) by:

$$X \in \{0, 1, 2, ...\}$$

whose support is given by $\Omega_X = \mathbb{N}_0$, $X(\omega) \in \mathbb{N}_0$.

Additionally, we may construct a R.V. for the number of wasps captured (during a direct-observation period of time) by:

$$Y_1 \in \{0, 1, 2, ...\}$$

However, the support of Y_1 is conditional on the maximum number of wasps visited. As such, we define the *conditional* support of Y_1 by $\Omega_{Y_1|X=x} = \{0, 1, ..., x\}$, $Y_1(\omega) \in \{0, 1, ..., x\}$, where x denotes the number of wasp visits during a (direct) observation period of time. However, the *marginal* support of Y_1 is of the form $\Omega_{Y_1} = \mathbb{N}_0$, $Y_1(\omega) \in \mathbb{N}_0$

Then, for the indirect observation portion, we may construct a R.V. for the number of wasps captured (during an indirect-observation period of time) by:

$$Y_2 \in \{0, 1, 2, ...\}$$

whose support is given by $\Omega_{Y_2} = \mathbb{N}_0$, $Y_2(\omega) \in \mathbb{N}_0$.

Note: Of course, we do not know the number of wasp visits during the indirect observation period.

Preliminary assessment about an assumption of independence:

The number of wasp visits and number of wasp captures, X and Y_1, are not independent, as the support of Y_1 necessarily depends on X. However, it seems reasonable (I'd argue) that Y_2 is independent of X and Y_1 given the difference in construction and design, i.e., the two different observation modes appear to cover non-overlapping periods of time, in plant-hours.