Lab 8 Handout – Multi-state models and time-varying covariates

The appropriate reference material to this lab are Chapter 12 in Powel and Gale and Chapter 10 in The Gentle Introduction, as well as the citations I've provided in the syllabus as supplemental references this week's discussion of multi-state models. Also see Blomberg et al. 2012. Journal of Avian Biology 44: 149-158 for more information on the data set and study system for Part I.

This week we'll be using a new analysis type, the multi-state, and will introduce a new parameter, the transition probability (Ψ). In addition to introducing you to this new analysis type, the goal today is to give you some tangible experience analyzing a slightly more complex dataset where we have to accommodate the modelling of multiple parameters, and analyzing both categorical effects and group and individual covariates that vary through time. First, a few words on multi-states, the data you'll be working with, and the questions we are interested in asking.

Multi-state models are a relatively simple extension of the CJS, where in addition to collecting data on captures and recaptures of individually marked animals, we also include information on some dynamic "state" variable of interest. States could consist of animal reproductive state (as we'll see today), physical location (e.g., presence in one or more study areas), nutritional status (good vs. poor body condition), or really any other plausible distinction you could think of. A key here is that membership in states is dynamic; animals have the ability to move among one or more states throughout their lifetime. However, while we have the ability to record that movement when we capture them, we also have an imperfect ability to observe state transitions when animals are not captured. The probability of movement among states is the transition probability (Ψ), which often can be related to interesting ecological questions. For example, if we are working on two adjacent populations of animals (State A and B) the value $\Psi_{A\text{-}B}$ (the transition from A to B) gives us an estimate of the rate of movement between A and B, and the value $1\text{-}\Psi_{A\text{-}B}$ gives us an estimate of the fidelity rate in population A, during any given interval. Thus multi-state models can be used in this context to evaluate population connectivity, and could contribute to better-understanding meta-population dynamics.

Including these state transitions also allow us to ask new and different questions about survival, and allows us to consider greater sources of heterogeneity in both survival and detection. Previously, we've incorporated similar sources of heterogeneity using group membership, but with the exception of age (really a state variable of sorts) we have assumed that group membership was fixed for each individual throughout their lifetime. By including a dynamic state-transition parameter, we can explicitly consider the instances where individuals change group membership among occasions. If survival or detection rate is influenced by state membership, we now have a tool for explicitly incorporating that variation. Consider the following conceptual model of the above scenario:

$$\Phi_{A1}\Psi_{A-A}$$

$$\Phi_{B1}(\Psi_{B-A})$$

$$\Phi_{A1}(1-\Psi_{A-A})$$

$$\Phi_{B1}(1-\Psi_{B-A})$$
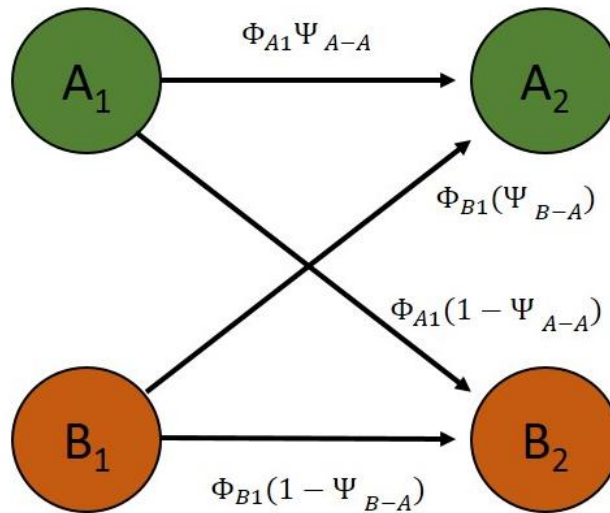
Here, the ability for an individual to "transition" from state (population) B to state A is a product of the survival probability of individuals in state B during interval 1 (ϕ_B), and the probability of transitioning from B to A between occasions 1 and 2 (Ψ_{B-A}). In contrast, the probability of remaining in state B is a product of the survival probability and one minus the B to A transition probability. Here there are only two options (A and B) so by default, the sum of the two transition probability values must equal one. Also note that there is a state-specific detection parameter not depicted above, which accounts for the observation process.

Importantly the state transitions are conditioned on having survived the previous interval, so these models inherently assume that state transitions happen instantaneously at the beginning of the each of occasion (t). This in turn implies an assumption that survival is conditioned on state membership in t-1; if the state transition occurred during the interval, and mortality occurred as a result of that unobservable state transition, we would typically have no way to record the transition and thus the mortality would be associated with the prior state. For this reason, carefully matching your sampling design (e.g. periods during which you capture animals) to the timing of state transitions (e.g. phenology of movements) is critical.

Hopefully the utility of including this additional state transition parameter is fairly intuitive. It improves our ability to accurately account for heterogeneity in our dataset, while also giving us the ability to ask new and interesting questions about the biology of our study organisms. In practice, the states are treated in Mark and RMark almost identically to a group structure with respect to setting PIMs and building the Design Matrix (MARK) or coding state effects in models (RMark). This means that we can test for support of state-level effects in our GLM/AIC hypothesis testing framework. Instead of the 0101 encounter history format we've seen so far, the file is coded as 0A0B, where the "1s" are replaced with a single character describing the state membership in each occasion. "0s" still reflect a lack of detection. Of course there are a number of

complexities and assumptions inherent to the multistate models that we need to be aware of. These are covered in substantial detail in the provided readings, and we'll some of them in greater detail when we cover the multi-state article during discussion.

Lab Exercise – Sage-grouse reproductive success.

Background: This week you will be conducting an analysis of reproductive success for greater sage-grouse and asking questions about how prior reproductive status, precipitation, and age affect the probability that a female sage-grouse successfully raises a brood. First, some relevant background on these data. This analysis was a component of one of my dissertation chapters, which dealt with survival and reproductive success of female greater sage-grouse in Nevada. In addition to asking general questions about temporal patterns in survival, we were also interested in evaluating whether costs associated with reproduction reduced the survival or future reproductive success of female sage-grouse. Our alternative hypothesis was that heterogeneity among females with respect to quality would override these reproductive costs, and we would find a positive association between current reproductive success and either future survival or reproductive output. We used data from 328 radio-collared female sage-grouse that were monitored from 2003-2011 in eastern Nevada. Of these, we collected data on reproductive success (success = raised 1 or more chicks to 45 days of age) for 272 females from 2005-2011, and these are the data you'll be using today.

Model specifications: For the larger survival analysis associated with these data, I used a known-fate analysis (see Ch 16 in the Gentle Introduction) to explore questions related to seasonal, annual, and age-associated variation in female survival, and asked whether reproductive status affected future survival. Because these were radio-marked individuals with mortality-sensitive transmitters, we could estimate survival in the known fate by assuming that detection = 1.0. However, the fact that the radio-collars had a life span of only ~ 3 years added a complication with respect to the reproductive analysis. This radio-failure added a potential source of heterogeneity to the data, because on average radios were more likely to fail for older rather than younger individuals. This meant we had a reduced ability to measure Ψ in older females. I realized that if I allowed the ϕ terms to estimate in a multi-state framework with the radio-failures included in the history as though they were unknown status (i.e. I did not right-censor those histories), it would affect the survival estimates in a similar manner as permanent emigration. This would have the effect of biasing the survival estimates low, as ϕ now reflected a product of both true mortality, permanent emigration, and radio failure. Including radio failure as a component of ϕ, however, effectively partitioned the potential variance associated with the radio failure into the survival component of the analysis, leaving the transition probabilities (presumably) free of this sampling bias. Thus in this analysis, the ϕ terms were effectively nuisance parameters, and because we were

concurrently evaluating survival using the known fate, the fact that the survival terms were substantially biased was not terribly problematic. This is a bit of a "non-traditional" application of the multi-state, but it also illustrates the flexibility of these methods.

In this analysis, the history consists of the "annual" reproductive status of females. The reproductive season for sage-grouse runs approximately between March (the onset of breeding) and July (the end of the brood-rearing period). We coded females as successful (state=S) if they both hatched a nest and had one or more chicks alive and present with them 45 days following the hatch of the nest. A female was classified as failed (state=F) when she either 1) was never found on a nest, 2) nested but lost the nest prior to it hatching, or 3) hatched the nest, but failed to successfully fledge the young.

Although detection was close to 1.0, we did have some instances where unique females were either not monitored over successive years, or were monitored during two distinct segments of the study with a gap of one or more years in between. It turned out by chance this was only true of unsuccessful females, so we should also evaluate possible state effects on both p and ϕ. The script provides guidance on how to model these nuisance parameters accordingly.

Time-varying covariates: This analysis includes two covariate vales that vary through time. The first is a group covariate for annual variation in precipitation within the system. It is referred to as a group time-varying covariate because the covariate values change through time, but during any time step are applied evenly across all members of a group. The values for the group covariate are as follows, which are z-standardized measures of annual rainfall for each year of the study.

Y1	1.90364
Y2	-0.85577
Y3	-0.52976
Y4	-0.02910
Y5	-0.59962
Y6	-0.05239
Y7	-0.28648

These data represent the total amount of precipitation falling as rain between August (year t-1) and July (year t) for each study year, as measured at a nearby weather recording station, and they are incorporated into the R script for the lab. For your information, the mean annual rainfall during this time period was 614 mm ± 289 SD (that's right, just greater than 6 cm of rainfall, or less than 3 inches, in an entire year … this is the desert, after all).

One important thing to remember is you should not mix group time-varying covariates (e.g. annual rainfall) with other similar temporal structures (time or year in this case) in the same model. This is because fully time-varying models already describe the full

range of annual variance in the data; fitting an additional temporal structure is redundant. In this case, a year + rain model would be nonsensical.

The second covariate we will consider is an individual time-varying covariate, where the value for the covariate changes through time in a manner that is (potentially) unique to each individual. For this analysis, we'll be looking at the effect of individual age, and we will use a time-varying covariate as a convenient tool to incorporate age effects. This is the first we've used a time varying covariate. The linear model takes the same form as a standard covariate model, with a single beta describing the covariate effect. In this case, a phi (State + Covariate) model would look like this:

Eq. 1 $$logit(\phi) = \beta_1 + \beta_2 S + \beta_3 X_i$$

And the matrix form of the model looks like this:

$$\begin{matrix} \phi_{A1} \\ \phi_{A2} \\ \phi_{A3} \\ \phi_{A4} \\ \phi_{B1} \\ \phi_{B2} \\ \phi_{B3} \\ \phi_{B4} \end{matrix} = \begin{bmatrix} 1 & 1 & X_1 \\ 1 & 1 & X_2 \\ 1 & 1 & X_3 \\ 1 & 1 & X_4 \\ 1 & 0 & X_1 \\ 1 & 0 & X_2 \\ 1 & 0 & X_3 \\ 1 & 0 & X_4 \end{bmatrix} * \begin{matrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{matrix}$$

Whereas in the past data for X were drawn from a single covariate value for each individual, in this case values are drawn from a sequence of covariates. These are incorporated into the data file as a sequence of covariate values (cov1, cov2, cov3, etc), and their implementation is described in the R script for the lab.

Estimating R²_Dev: When using a time-varying group covariate, we can generate an approximation of the proportional variance in a demographic rate that is explained by the variable using an approach originally presented by Grosbois et al. (2008: Biological Reviews 83:357-399), which is given as

Eq. 2 $$R^2_{Dev} = \frac{Dev(F_{dot})-(F_{cov})}{Dev(F_{dot})-(F_{time})}$$

Here we compute our r-squared approximations based on the model deviances, which are returned in the model selection table for each model set. F_dot refers to the null model (or its equivalent; read on), F_cov the model that contains a temporal covariate (e.g. precip in our case) and F_time is a model that is fully time-varying (e.g. the year model). The great leap of faith here is that if we compare the ratio of distance in model

deviance between a covariate model and one with no temporal variation (the numerator) to the distance in model deviance between a fully time-varying model (that explains all temporal variance) and a constant model (the denominator), that should approximate the proportion of the full temporal variance explained by the covariate. For this reason, the F_dot model need not be the literal null model, it could be a base model with no temporal variation (e.g. a Psi(state) model) so long as the F_cov and F_time models also contain the same basal structure (i.e. Psi(state + cov) and Psi(state + time), respectively).

R2_Dev is a very handy tool for computing what many refer to as a pseudo-r-squared, that allows you to not only attribute significance and slope of a covariate effect, but also describe how important that affect is in explaining the overall variability of the system. For that reason I recommend incorporating this into your analyses if possible as it strengthens your interpretation of the temporal dynamics of your system.

If you are really savvy you can also extend these concepts to spatial variation in some cases. For example, if you have discrete study areas that vary by a measurable covariate (e.g. habitat composition) you could ask how much of the among-area difference is attributable to the habitat variable. See Gibson et al. 2014 (Ibis 156:729-740) for an example application. This approach is not appropriate for use with individual covariates.

Darrell Pruett

Darrell Pruett
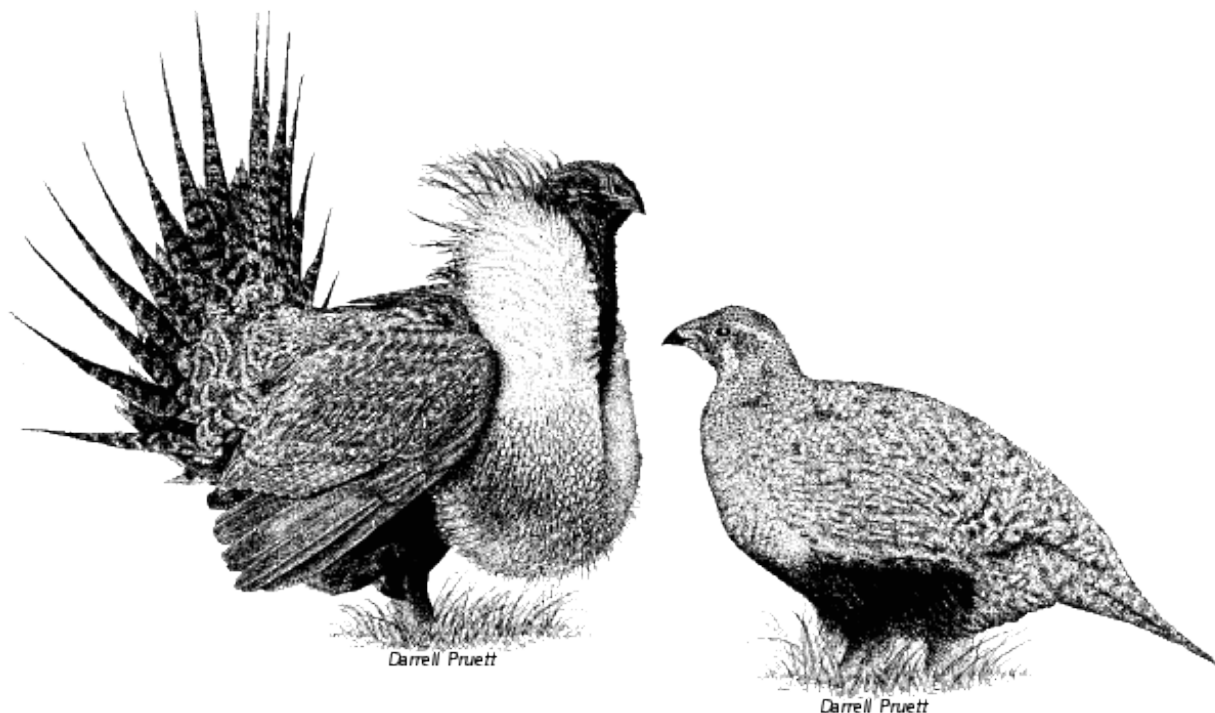
Lab 8 Assignment – Elk in the Beaver State

For this week's assignment you'll be analyzing a real dataset that comes from North American elk monitored at the Starkey Experimental Forest in eastern Oregon. Starkey contains a very large (10,000+ acre) enclosure that was originally designed to study the effects of forest practices on populations of elk and mule deer, where the density of the ungulates could be experimentally manipulated (because of the enclosure). Each winter, the elk herd is rounded up on winter feed grounds, animals are marked, and an ultrasound is used to measure the pregnancy status of the cows and their winter body condition based on the maximum depth of their rump fat (MaxFat; MF). This data was used by Sabrina Morano (Morano et al. 2013. J. Mammalogy 94:162-172) to evaluate factors that cause variation in the survival and pregnancy rates of female elk, where the "factors" can include things like age (yearling vs. adult), individual body condition (based on MaxFat), wintertime precipitation, previous year's reproductive status (state, Pregnant or Not Pregnant), and general temporal variation. It turns out these questions are perfectly suited for a multistate model.

I would like you to analyze Sabrina's data to address the following questions:

1. Does previous reproductive status affect current reproductive status?
2. Does reproductive status affect survival?
3. Does body condition influence either reproduction or survival?
4. Does winter precipitation influence either reproduction or survival?

On the assignments page of Blackboard you will find an input file named Elk.csv that contains the data. Notice that I've removed a few of the covariates that Sabrina used in her analysis, as well as the age structure, to simplify the assignment a bit. So, while you can use her paper as a check on the results you find, you shouldn't necessarily expect to reach identical conclusions to what is reported in the paper. Below are the winter precip values (z-standardized). These data also have fewer peculiarities compared to the sage-grouse data, so you will need to decide which pieces of code you need to use, or not, compared to the sage-grouse analysis.

Year	Winter Precip
Y1	-1.36
Y2	0.03
Y3	1.24
Y4	0.89
Y5	-0.88