Homework 8: How the Recent Mammals Got Their Size Distribution

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 1 April 2013

Homeworks 4 and 6 used regression to study how the typical mass of (mammalian) species changes over evolution: on average new species are heavier than their ancestors, especially if the ancestor was very small, but with a wide variation. If we combine this with the facts that new species branch off from old ones, and that sometimes species go extinct without leaving descendants, we get a model for how the distribution of body masses changes over time. It's not feasible to say much about this model mathematically, but we can simulate it, and check the simulated distribution against the real distribution of body masses today.

The objects in this model are species, each described by its typical mass. (We assume that this does not change over the lifespan of the species.) Each species can produce new species, who mass is related to that of its ancestor according to our previously-learned regression model, or go extinct. As time goes on, the distribution of body masses will fluctuate randomly, but should do so around a steady, characteristic distribution.

More specifically, each species i has a mass X_i, which is required to be between $x_{\min}$, the smallest possible mass for a mammal, and $x_{\max}$, the largest possible mass. At each point in time, one current species A is uniformly selected to evolve into exactly two new species. Each descendant has a mass X_D which depends on the mass of its ancestor, X_A, according to the regression model, plus independent noise:

$$\log X_D = \log X_A + Z + \left\{ \begin{array}{cl} a + b \log X_A & \text{if } \log X_A \leq d \\ c & \text{if } \log X_A \geq d \end{array} \right. \quad (1)$$

where $Z \sim \mathcal{N}(0, \sigma^2)$. Continuity means that $a = c - bd$; we also need to impose the constraints that $x_{\min} \leq X_D \leq x_{\max}$.

Species become extinct with a probability that depends on their body mass,

$$p_e(x) = \beta x^{\rho} \quad (2)$$

Unless otherwise specified, you should use $\sigma^2 = 0.63$; $x_{\min} = 1.8$ grams and $x_{\max} = 10^{15}$ grams; $\rho = 0.025$; $\beta = 1/5000$; and the values of b, c and d from the solutions to Homework 4.

1. (10) Write a function, `rdeac.1`, which takes as inputs a single ancestral mass X_A (not $\log X_A$), the parameters$b$, c, d and σ^2, and the limits $x_{\min}$ and $x_{\max}$. It should generate a candidate value for X_D (not $\log X_D$) from Eq. 1 and return it if it is between the limits, otherwise it should discard the candidate value and try again.

   (a) (2) Set X_A to 40 grams and check, by simulating many times, that the output is always between $x_{\min}$ and $x_{\max}$, even when those values are brought close to 40 grams.

   (b) (8) Simulate a single X_D value for 100 values of X_A evenly spaced between 1 and 100 grams. Treat this as real data and re-estimate the parameters b, c and d according to the methods of Homework 4; are they reasonably close to those in the simulation?

2. (10) Write a function, `rdeac`, which takes the same inputs as `rdeac.1` *plus* an integer n, and returns a vector containing n independent draws from this distribution. We will test this with $n = 2$, but your code must be more general for full credit.

   (a) (4) Check, by simulating, that the first component of the returned vector has the same marginal distribution as the output of `rdeac.1`.

   (b) (4) Check that the second component of the returned vector has the same marginal distribution as the first component.

   (c) (2) Check that the two components are uncorrelated.

3. (10) Write a function, `speciate`, which takes the same arguments as `rdeac.1`, except that X_A is replaced by a vector of ancestral masses. The function should select one entry from the vector to be X_A, and generate two independent values of X_D from it. One of these should replace the entry for X_A, and the other should be added to the end of the vector.

   (a) (2) Check, by simulating, the output always has one more entry than the input vector of masses, no matter how long the input is.

   (b) (8) If the input has length n, check that $n - 1$ of the entries in the output match the input.

4. (15) Write a function, `extinct.probs`, which takes as inputs a vector of species masses, an exponent ρ, and a baseline-rate β, and returns the extinction probability for each species, according to Eq. 2.

   (a) (1) Check that if the input masses are 2 grams and 2500 grams, with the default parameters the output probabilities $\approx 2.0 \times 10^{-4}$ and 2.4×10^{-4} respectively.

   (b) (2) Check that if $\rho = 0$, then the output probabilities are always β, no matter what the masses are.

(c) (2) Check that if there input masses are all equal, then the output probabilities are all the same, no matter what ρ and β are.

(d) (10) Write a function, `extinction`, which takes a vector of species masses, ρ and β, and returns a possibly-shorter vector which removes the masses of species which have been selected for extinction. *Hint:* What does `rbinom(n,size=1,prob=p)` do when `p` is a vector of length `n`?

5. (15) *Evolve!*

   (a) (5) Write a function, `evolve.1`, which takes as inputs a vector of species masses, b, c, d, σ^2, $x_{\min}$, $x_{\max}$, ρ and β, and first does one speciation step, then one round of extinction, and returns the resulting vector of species masses.

   (b) (5) Write a function, `evolve`, which takes the same inputs at `evolve.1`, plus an integer t, and iterates `evolve.1` t times.

   (c) (5) How do you know that your functions are working properly?

6. (15) *Re-running history*

   (a) (5) Run `evolve` starting from a single species with a mass of 40 grams for $t = 2 \times 10^5$ steps. Save the output vector of species masses as `y1`. Plot the density of `y1`.

   (b) (5) Repeat the last step to get a different vector `y2`. Does it have the same distribution as `y1`? How can you tell?

   (c) (5) Change the initial mass to 1000 grams and get a vector of final masses `y3`. How does its distribution differ from that of `y1`?

7. (25) The data file MOM_data_full.txt gives the masses of a large (and representative) sample of currently-living species of mammals. The column `mass` gives the mass in grams; the columns `species`, `genus`, `family`, `order` and `code` are identifiers for the particular species, which do not matter to us. Finally, the column `land` is 1 for species which live on land and 0 for those which live in the water.

   (a) (5) Load the data and plot the density of masses for land species.

   (b) (10) Describe, in words, how the distribution of current species masses compares to that produced by the simulation model in `y1`.

   (c) (10) Use the relative distribution method from Chapter 16 to compare the actual distribution to the distribution of `y1`. Describe the results and what they say about how the data differ from the model.