7.7 Phylogenetic Regression with a Mixture of Discrete and Continuous Data in R

Introduction

Compared to analyzing regressions among continuously distributed data, a more complex situation arises when the dependent variable is binary while the predictor (independent) variables are continuous or are a mixture of binary and continuous data. For example, suppose we wanted to use body size to predict whether females of a primate species would exhibit sexual swellings, and we wanted to do so in a phylogenetically informed way? We could hypothesize that smaller females might have less energy surplus, and so would be less likely to evolve sexual swellings.

Sexual swellings are usually scored for any species as present or absent. Because there are only 2 states of the dependent variable, we cannot use a least square regression because there is a nonlinear relationship between values of the independent and dependent variables. In the case of a binary dependent variable, this nonlinearity results from their being a single floor (0) and ceiling (1) for the trait such that a sigmoidal relationship exists between the binary variable and the predictive independent variable.

We can solve this problem by conducting a binomial regression model (also called a logistic or logit regression model) with a correlation or covariance structure that expresses the nonindependence of phylogenetically related data. The exact implementations to do this are still under development, but the fundamentals of phylogenetic logistic regression are no different from phylogenetic general least squares regression. Both use a correlation or covariance matrix to transform the data points to accommodate their nonindependence (technically the correlation/covariance of the residuals in the model; Rohlf 2006). By doing this transformation, it is possible in theory to run any form of generalized linear model, even ones that apply to non-continuous dependent variables (Paradis and Claude, 2002).

Running the model in R

Here I will show how to run a phylogenetic logistic regression of sexual swellings on female body size using the 'ape' package in R. To run the analysis in R, first download the data file.

SwellingMassOWM.txt

This file contains data for a number of Old World monkeys (Cercopithecidae). The first column contains the species names, the second column contains log female body mass, and the last column contains the sexual swelling state. I have already log transformed female body mass because body mass frequently exhibits a large upper tail that causes it to deviate from normality (incidentally this type of departure from the Gaussian 'normal' distribution is called skew). This log transformation would be done in a least squares regression of continuous data as well - it has nothing to do specifically with the logistic regression model we will be using.

In R, navigate to the folder where you put the data file. Load the file into R by typing the following:

```
swelling.data = read.table("SwellingMassOWM.txt",row.names=1)
```

The file is now loaded in to R as the object 'swelling.data'. Now we need a tree to use during our logistic regression. Download the tree file.

TreeSwellingsMass.nex

Then, put it in the same folder where you have the data file. Load 'ape' and the tree into R:

```
library("ape")
```

```
tree = read.nexus("TreeSwellingsMass.nex")
plot("tree")
```

Notice that the names in the tree and the data table match exactly. The names have to match exactly. You can also run the test without any names in the data table, but this will assume that the species are ordered by row as they are numbered in the tree. It is best to use the names and make them match rather than chance having a row out of order.

Now we are ready to run a phylogenetic logistic regression. We will use the function, 'compar.gee' in the 'ape' package. This function models the nonindependence effect of phylogeny as a correlation structure among the tip data. It implements the methods described in Paradis and Claude (2002). Type the following:

```
logit.model = compar.gee(Swelling~FemaleWeights,data=swelling.data,family="binomial",phy=tree)
```

This single line of code fits a logistic (a.k.a. logit) model with the phylogeny expressing the expected nonindependence of the species data points. The 'compar.gee' calls this function, while the material in the parentheses tells the function information it needs. Thus, 'Swelling~FemaleWeights' tells the function that we are regressing sexual swelling status on female body weight. The code 'data=swelling.data' tells the function the data frame that contains the variables 'Swelling' and 'FemaleWeights'. The family argument is set to 'binomial', which means we are running a logit model (as opposed to one of many other regression models). The phy argument says the relevant phylogeny is named 'tree'.

Now type 'logit.model' to see the results of the fitted model. You should see the following:

The slope estimate (0.039) shown above indicates that the presence of sexual swellings is associated with larger female body mass, which was as predicted. This association, however, does not appear to be significant, as the two-tailed p-value of the slope is 0.72 (this is shown in the row 'FemaleWeights' under the header 'Pr(T > |t|)'). Is this lack of significance caused by the phylogeny?

When the phylogeny accurately describes the nonindependence of data points, then using the phylogeny actually increases statistical power. That is, using a phylogeny makes us more likely to find an association when there truly is one. Using the phylogeny also reduces type 1 error of inferring an association when none exists (Rohlf 2006). All this is dependent, however, on the first statement of this paragraph: that the phylogeny accurately describes the nonindependence. If this is not true, then the phylogeny can make things worse statistically.

The best way to assess how the phylogeny should be incorporated into the statistical model is to use a scaling parameter that adjusts the importance of the phylogeny. The value of the scaling parameter is usually chosen on likelihood or Bayesian grounds (see Chapter 5). Ives and Garland (2010) present a convincing argument that some form of scaling for phylogenetic signal should actually always be used in a logistic regression. This is because a binary dependent variable (unlike a continuous one) has a highly bounded distribution such that phylogenetic signal decays over time even under random Brownian motion. Put another way, for a binary character the time-dated branch lengths of a phylogeny do not predict character divergence under Brownian motion, even though these branch lengths are good predictions of continuous characters that evolve under the same model.

Phylogenetic logistic regression models that use scaling parameters are not yet developed in R, unlike the case for phylogenetic generalized least squares models for which R can implement a variety of scaling parameters (see AnthroTree 5.7 and other examples for Chapter 5).

You can run the logistic regression without a phylogeny easily in R with the following code:

```
logit.model = glm(Swelling~FemaleWeights,data=data,family="binomial")
summary(logit.model)
```

Notice how in the model output the effect size of female body mass on sexual swellings is an order of magnitude larger than in our phylogenetic model, and that the associated p-value is much smaller.

Note: One program is available that implements a scaling parameter for phylogenetic signal concurrently in a phylogenetic logistic regression. This model runs in MATLAB, and the code is available by emailing Ted Garland. I used the MATLAB model to examine the swellings data from above. The results indicated that the phylogeny did not reflect the nonindependence of data point residuals in the model very well (i.e., the phylogenetic signal was low). There also was a significant positive association of sexual swelling presence and female body mass. So, in this particular case, using the phylogeny without the scaling parameters actually appears to be incorrect.

References

> *Ives, A.R. and T. Garland Jr. 2010. Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59: 9-26.*

> *Paradis, E., and J. Claude. 2002. Analysis of comparative data using generalized estimating equations. Journal of Theoretical Biology 218:175-185.*

> *Rohlf, F. J. 2006. A comment on phylogenetic correction. Evolution 60:1509-1515.*

Contributed by Luke Matthews