This is a video transcript of Module 6 of the Online Open Course in Species Distribution Modelling. To access the full suite of videos click here.

Online Open Course - Species Distribution Modelling

Powered by the Biodiversity and Climate Change Virtual Laboratory

Module 6 Statistical regression models

Welcome back to this online open course about species distribution modelling. In the previous module, we looked at models for which you only need to provide occurrence data to predict the distribution of a species. In this module, we will focus on statistical regression models, which use both presence and absence data.

We covered in module 3 that absence data can either be true absence, or if such data is not available you can 'make up' absence data, which is referred to as pseudo-absence data. Presence-absence algorithms compare the environmental conditions of occurrence sites with those of absence sites.

The statistical regression models that I will explain in this module use all the data available to estimate the coefficients of the predictor variables, and they construct a function that best describes the effect of the environmental variables on species occurrence. The suitability of a particular model is often defined by specific model assumptions. In this module, I will explain the background of three statistical regression models: generalized linear models, generalized additive models and the multivariate adaptive regression splines technique. These models are an extension of 'simple' linear regression models.

Linear regression models work on a few assumptions, such as the assumption that we can use a straight line to describe the relationship between the response and the predictor variables. This implies that a constant change in a predictor leads to a constant change in the response variable. For example, the number of cricket chirps increases with increasing temperature. In a linear regression model you would assume that an increase of 1 unit in temperature will lead to an increase of 1 unit in the number of chirps. This assumption is often violated in species distribution modelling, particularly when the response is a probability, like the chance that a species is present at a given location. In this case, a generalized linear model is used, which allows more flexibility in the distribution of the response.

Generalized Linear Models (GLM)

A generalized linear model with binomial data, such as the presence and absence of a species, is called a logistic regression. If we would draw a line through these data points, it would look like an S-shape, and this line represents the probability of species occurrence. So, if all your species data points are absences, the probability is 0, whereas if all your points are presences, the probability is 1. At the value of the environmental variable where absence changes into presence, the corresponding probability is 0.5. This means there is an equal number of occurrence points and absence points recorded at that value of the environmental variable.
But of course, in a species distribution model, we are not looking at the effect of only 1 environmental variable, but at several different variables. In a generalized linear model, all predictors that you put into the model, are combined in an overall score for the environmental suitability, the so-called ‘linear predictor’.

Because the probability of species occurrence is based on binomial data that are not normally distributed, we need a function that links the response variable to the linear predictor. This is a so-called link function, and for binomial data this is a logit function. With the logit function, we take the log of the odds ratio, which is the ratio of the probability of presence over the probability of absence. This link function transforms the y-axis to be able to fit a straight line between the response and the linear predictor. The formula that describes this line operates on the log odds scale, and has a constant baseline, which refers to the logs odds ratio that you get when you set all environmental variables to their mean values, and it is the intercept of the straight line with the y-axis. And then the effect of each environmental variable is included by multiplying the value of the variable with its coefficient, and this affects the slope of the line.

Generalized Additive Models (GAM)

The next statistical regression algorithm that I’ll explain is the Generalized Additive Model, or GAM. Generalized additive models are an extension of generalized linear models, and have the same features that are unique to these types of models: the linear predictor, that combines all environmental variables that you put into the model in an overall environmental suitability score, and the link function that transforms the response into the log of the odds ratio. While GLMs construct a linear function between the response and the predictors, a GAM takes into account that the relationship might be of a more complex form and not entirely straight. To fit this complex relationship, the coefficients of the predictor variables in the linear predictor are replaced by a smoothing function. For each of the environmental variable in the model, the GAM algorithm calculates a smooth function that fits the data as closely as possible. In the final model, the smooth functions of each of these variables are added. Because GAMs are additive, it is difficult to include interactions between predictors and this is thus not often done. There are a lot of different smoothing functions available, but in species distribution models the most widely used is the cubic smoothing splines method.

Multivariate Adaptive Regression Splines (MARS)

The last statistical regression algorithm that I’ll explain in this module is the Multivariate Adaptive Regression Splines algorithm. While we group this algorithm within the category of the statistical regression models, it also has characteristics that are similar to machine learning models. The Multivariate Adaptive Regression Splines algorithm is another extension of linear models and it is a powerful algorithm that is able to model complex relationship between the response variable and the predictor variables.

I'll explain how the algorithm works with this example in which we plot the probability of occurrence for a species against an environmental variable. It is clear that these points do not reflect a linear relationship between the predictor and the response. The Multivariate Adaptive Regression Splines algorithm divides the range of predictor values in several groups and for each group, a separate linear regression is modeled, each with its own slope and its own associated error. The separate lines are connected, and these connections are called knots. The Multivariate Adaptive Regression Splines algorithm automatically searches for the best spots to place the knots. Each knot has a pair of basis functions. These basis functions describe the relationship between the environmental variable and the response. The first basis function takes the maximum value out of two options: it is either 0 or the result of the equation 'Env var value – Env value of knot'. So in this example the knot is at the value 11 for the environmental variable. For any value below 11, the outcome of the equation 'Env var value – Knot' will result in a negative number, which is smaller than 0 and thus the outcome of the basis function is 0. This means that the outcome of basis function 1 is 0 for all environmental values up to the knot, while for all values after the knot the outcome of basis function 1 is the value of the environmental variable minus 11. The second basis function has the opposite form, with the outcome of 0 for all environmental values after the knot, and the outcome of 11 minus the value of the environmental variable before the knot.

For a simple model with only 1 environmental variable and only 1 knot, the final model includes a baseline, which is the mean of the response values, and the two basis functions on each side of the knot. However, most models will include multiple knots and multiple environmental variables, leading to more complex models.

In general, these three statistical regression models are very useful as they can all handle both continuous and categorical predictor variables and they are designed to fit complex relationships between the predictors and the response. Generalized Additive Models and Multivariate Adaptive Regression Splines are more robust to outliers than GLMs, but in return more susceptible to overfitting. Generalized Linear Models and Generalized Additive Models are able to work well with smaller datasets, but it is good to keep in mind that with the more predictor variables you want to include in the model, the larger your occurrence dataset needs to be. As a rule of thumb, the number of predictor variables should be less than the total number of occurrences divided by 10. So, if you have an occurrence set of 100 points, you want to include a maximum of 10 environmental variables as predictors in your model. Multivariate Adaptive Regression Splines can handle large datasets very well, and despite the complexity of this algorithm it tends to work very efficiently and fast. While these algorithms are a bit more complex to understand and interpret, they are definitely worth considering when you want to run a species distribution model.

Thank you for watching module 6 of the species distribution modelling course. In module 7, we will go into the details of the last group of models, the machine learning models.

Attribution

Please cite this video as follows:

Huijbers CM, Richmond SJ, Low-Choy SJ, Laffan SW, Hallgren W, Holewa H (2016) SDM Online Open Course, module 6: statistical regression models. Biodiversity and Climate Change Virtual Laboratory, http://www.bccvl.org.au/training/. DDMMYY of access.

Acknowledgements

Lead partners: Griffith University, James Cook University

Thanks to: University of New South Wales, Macquarie University, University of Canberra

Funded by: National eResearch Tools and Resources Project (NeCTAR)

The BCCVL is supported by the National eResearch Tools and Resources Project (NeCTAR), an initiative of the Commonwealth being conducted as part of the Super Science Initiative and financed from the Education Investment Fund, Department of Education. The University of Melbourne is the lead agent for the delivery of the NeCTAR project and Griffith University is the sub-contractor.