De : Cade, Brian S <cadeb@usgs.gov>
Envoyé : mercredi 13 avril 2022 17:09
À : Amélie Lehuen <amelie.lehuen@unicaen.fr>
Cc : Barry.Noon@colostate.edu
Objet : Re: [EXTERNAL] RE: Quantile regression advice for species distribution model

Amélie:

As some follow-up to my comments yesterday, I am including some example code and a graph in the attached Word document. I simulated some data using your Gaussian distribution as the response form (1 predictor) but then estimated the 0.95 and 0.99 quantiles using linear quantile regression (rq) with cubic b-splines.

You can see in the graph in the Word file that the cubic b-splines with a knot at median of the predictor capture the functional form very well. Of course, the advantage of the cubic b-splines over explicitly specifying the Gaussian model in nlrq() is that most data will never be as explicitly limited by a Gaussian equation as these simulated data.

Most will likely be shifted asymmetrically around the median and the nonlinearity of the curve is unlikely to always take the exponential form of the Gaussian. Here, the cubic b-splines would recover the form more reasonably than trying to force the analysis into fitting into the more restrictive model shape done when you specify the Gaussian model form explicitly in nlrq() model.

Hope this helps you see the advantages of potentially fitting your nonlinear biomass response with a model that is more flexible than the explicit specification of the Gaussian form. Of course, you could estimate both and compare. But my guess is the flexible b-spline approach will provide better fit most of the time.

Brian

Brian S. Cade, PhD
U. S. Geological Survey
Fort Collins Science Center
2150 Centre Ave., Bldg. C
Fort Collins, CO 80526-8818
email: cadeb@usgs.gov
tel: 970 226-9326

De : Cade, Brian S <cadeb@usgs.gov>
Envoyé : mardi 12 avril 2022 19:48
À : Amélie Lehuen <amelie.lehuen@unicaen.fr>
Cc : Barry.Noon@colostate.edu
Objet : Re: [EXTERNAL] RE: Quantile regression advice for species distribution model

Amélie: Okay, great to see the additional information so that I now can see what you meant by a Gaussian function. I have a couple of suggestions that might allow for simpler modeling by taking estimation into linear quantile regression with rq() rather than by estimation using nonlinear nlrq(). The nonlinear nlrq() estimation approach is more sensitive to data vagaries and the algorithm is not as well behaved as the linear rq() algorithm, nor are there as many inferential and utility functions available for nlrq() estimates as you are discovering.

I think you are interpreting the ecological literature a little too literally when interpreting your response functions as Gaussian. Certainly there are many examples where there is a nonlinear function, increasing to some peak value and then decreasing. While this has often been described as a Gaussian functional form, there is no reason to think that the curve has to take on the specific mathematical form of a Normal distribution.

The key functional form properties are a nonlinear increase to some peak followed by a nonlinear decrease. You could easily avoid the Normal distribution functional form that you are trying to specify in nlrq() with something equivalent but with a more flexible function form by using splines in linear rq() function. For example, if y is your biomass response variable, x1 is your bed shear stress predictor variable, and x2 is your percent mud content predictor variable, you could estimate the following model with b-splines in linear quantile regression:

```
   library(quantreg)
   library(splines)
   med.x1 <- median(data$x1)
   med.x2 <- median(data$x2)

   model.1 <- rq(y ~ bs(x1,degree=3,knots= med.x1) + bs(x2,degree=3,knots=
med.x2),data=data,tau = c(0.50,0.75,0.80,0.90,0.95))
   summary(model.1,se="boot",bsmethod="wxy",R=1000)
```

The degree=3 argument in b-spline function bs() fits cubic polynomials using basis specification, and the knot argument identifies a breakpoint where the curve can change. Here I set it as the median of the predictor variable value, but you can experiment with alternative values that might provide a better fit. For example, you can compare model estimates with different values for the knots by using AIC or similar information criterion.

You can force the knot locations to be identical for different quantiles as above, or estimate different quantiles separately with different knot locations specified for different quantiles (tau). So this model.1 specification would allow the response surface to have a nonlinear functional form that increases to a peak near the knot value provided and then to have a nonlinear decrease at higher values of the predictors, if the data supports such a response shape. My specification above forces the knot locations to be identical for each quantile, but the nonlinear spline response surfaces can differ among quantiles otherwise.

Other response functions (e.g., closer to linear) can also be estimated with the same model form. As with your nlrq() modeling, the nonlinear response forms of the two predictors x1 and x2 are allowed to take different shapes.

The advantage of estimating this nonlinear response surface with linear rq() estimator is that you have access to a much wider range of inferential procedures (e.g., bootstrapping), and utility functions for prediction and graphing, and a much more stable estimation algorithm than the nlrq() estimation approach. The summary() statement I provided uses one of the bootstrap procedures appropriate for heterogeneous data with R = 1000 resamples. There are many other alternatives. You will get the quantile response surfaces by using the predict(model.1,newdata=newdata) function with appropriate range of values for predictors x1 and x2 given in newdata.

You might try this linear quantile regression approach with b-splines and see how it compares to your nlrq() estimates using Gaussian response functions. If the Gaussian response functions in nlrq() were good fits to the data, then I would expect the response surfaces to be fairly similar between the two approaches. If the Gaussian response functions are not very good fits to the data, then estimates between the nlrq() and rq(. bs()) approaches might differ a fair bit because the latter is far more flexible.

Another possibility with splines to estimate nonlinear functions similar to above would be to use qgam package. This quantile regression estimator can fit similar nonlinear spline (I think it uses thin plate splines by default) functions but where the estimation process tries to optimize where the knots are located rather than by you having to specify them in bs() function within rq(). I am not usually bothered by having to select knots in bs() specification so I do not usually make use of qgam but it is there as an option.

With the sample sizes you have, it is probably unreasonable to expect estimates for very high quantiles (tau > 0.95) to be estimated very reliably, but this is partly dependent on actual data patterns. The cubic b-spline functions on the two predictors, x1 and x2, would require 8 degrees of freedom (df).

I hope this helps point you towards some useful alternatives. Sorry to hear about you catching COVID. Hope you have recovered well.

Brian

From: Amélie Lehuen <amelie.lehuen@unicaen.fr>
Sent: Tuesday, April 12, 2022 8:54 AM
To: Cade, Brian S <cadeb@usgs.gov>
Cc: Barry.Noon@colostate.edu <Barry.Noon@colostate.edu>
Subject: [EXTERNAL] RE: Quantile regression advice for species distribution model

Hi Brian,

Thanks for your quick answer, I was not expecting that. I am very sorry to make you confused, I forgot to add the abstract named 'MP SDM-NEO', the reason why you did not found anything to enlighten you.

I finally got caught by the COVID, so it has taken me longer to come back to you. I have made some support to explain what I am aiming, in a form of a very quick summary you will find enclosed added to the abstract, that could help. I enclosed also the abstract of the project MELTING POTES, to contextualize.

We basically have the same kind of data set as Cozzoli et al. (2014) in the Seine estuary in France, and want to model the distribution of selected species. The global idea here is to keep a 'biological process' approach, meaning that a mathematical solution can work, but may be difficult to export in another estuary, which is part of our goal. We want also to use the quantile regression for its interest with long time series data, to discard exceptional events that may reduce the biological response. So, based on the fact that the biological response is frequently following a gaussian law, we figured that we could use the gaussian law with two factors, in order to increase the accuracy of the model.

I made that work with the quantreg package in R and nlrq function, with the equation you can find in the Summary. I had also made the one factor gaussian and the linear quantile regression for comparison, but overall the feeling is the same. My main concerns are:

The result of the nlrq can have big variations from a quantile to another, as much as having some qauntile that are ok, and other that are totally wrong (see examples in summary)

I have not a model validation process per say, visual evaluation being the main tool I use. I use the AIC for the selection of factors among the 6 different I have, even though a visual check is really revealing that sometimes it does not goes as expected with a 'good' AIC

The proper use of the nlrq function with 6 different species, where n goes from 347 to 1152. I have chosen to use the "BFGS" method for its worked better, but I really don't know what is contains (didn't dare try the "L-BFGS-B" which may be useful to keep solution with relevant values). The nlrq command does not seem to have a bootstrap method as Cozzoli used in the rq, the linear regression (which I am not sure I totally got).

You adequately mentioned the model validation, because that seems to me the heart of mastering the quantile regression. My best AIC is not my best model if I consider all the quantile I test. I have chosen to keep the models that are looking best on several quantile rather that the one that have a only one good quantile result.

I have explored the mathematical papers on the subject, but I am in my limits to have the felling of getting it, and thus, use it correctly!

I hope I made my case clear enough. I you want we can settle a meeting to exchange live, but it has to be before May, 11th, because I am starting an experimental phase that will be full time for the next 3 months.

Thanks again for your help, it means a lot,
Amélie Lehuen - +33 2 31 56 51 02 - +33 6 72 18 94 51

De : Cade, Brian S <cadeb@usgs.gov>
Envoyé : mardi 29 mars 2022 18:43
À : amelie.lehuen@unicaen.fr
Cc : Barry.Noon@colostate.edu
Objet : Quantile regression advice for species distribution model

Amélie: I would be happy to provide you with some advice on using quantile regression for species distribution modeling. I have read the Cozzoli et al. (2014) as well as earlier (2013) papers on quantile regression for species distribution modeling. There certainly are some issues with model validation that might be done better than what they did, e.g., using quantile conformal predictions with a withheld validation data set rather than their sampling without replacement strategy. I am not sure what you are implying with your "bivariate gaussian equation see 'MP SDM-NEO' for details." I could not find MP SDM-NEO so I do not know what this refers to. You will need to provide me with more information on this approach for me to be able to comment usefully.

There are, of course, many alternatives for getting nonlinear smoothing functions of predictors into quantile regression. B-splines are easy to implement on a predictor by predictor basis. The quantreg package in R also offers some bivariate smoothing spline options which often are useful for spatial coordinates (latitude and longitude). There also now is an automated generalized additive model for nonlinear smoothing available in the qgam package for R.

So there are many potential things we can discuss. Looking forward to hearing from you.

Brian

From: Amélie Lehuen <amelie.lehuen@unicaen.fr>
Sent: Thursday, March 24, 2022 3:53 AM
To: cadeb@usgs.gov; Noon,Barry <Barry.Noon@colostate.edu>
Cc: 'Francis Orvain' <francis.orvain@unicaen.fr>
Subject: Quantile regression advices for species distribution model

Dear M. Cade and M. Noon,

I am working as a PhD student in Normandy, France on estuaries ecosystem modelization, in the project MELTING POTES. My director Francis Orvain, and I are working with some partners in Netherland and Italy, in particular Francesco Cozzoli, that have published work on the use of quantile regression to create species distribution models :

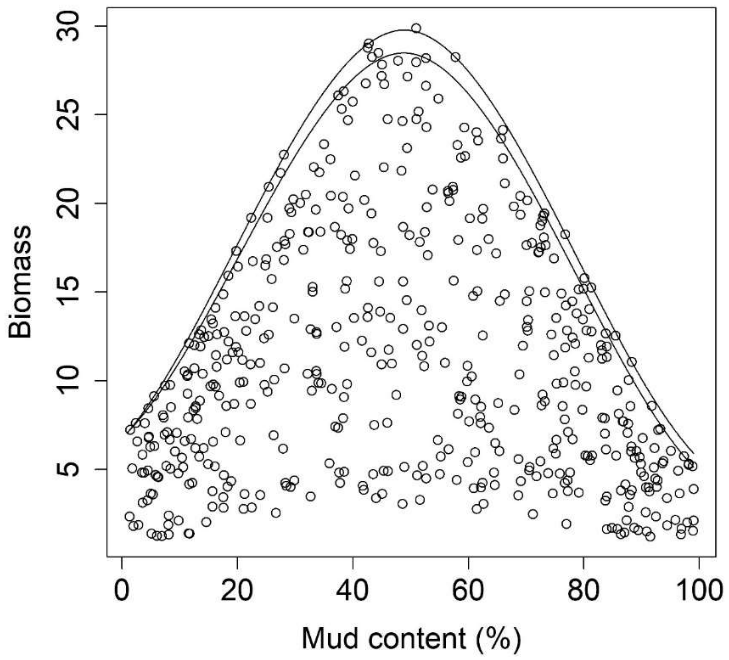
Cozzoli, F., Eelkema, M., Bouma, T.J., Ysebaert, T., Escaravage, V., Herman, P.M.J., 2014. A Mixed Modeling Approach to Predict the Effect of Environmental Modification on Species Distributions. PLoS ONE 9, e89131. https://doi.org/10.1371/journal.pone.0089131

We are building on that to go further, and I have developed a model that use quantile regression in non-linear, with a bivariate gaussian equation (see 'MP SDM-NEO' for details). I have read as much as my non mathematical trained brain allowed me, but I have still concerns to use it correctly. We are really convinced of the relevance of this method for biological studies, so it is time to have exchanges with experts in that domain.

So if you are interested, and available of course, we can organise something at your convenience, or if you have people to recommend to exchange with, we would be more than glad.

Have a good day,
Best regards,
Amélie Lehuen - +33 2 31 56 51 02 - +33 6 72 18 94 51

```
### Example generation of ecological response data with limiting equation
### specified explicitly by a Gaussian form, but then estimated with cubic
### b-splines in quantile regression.

### Here sample size is 500, x is constructed like percent mud content (1 -
### 99%.

x <- runif(500,min=1,max=99)
med.x <- median(x)
sd.x <- sd(x)

y <- 30*(exp(-1*(((x - med.x)^2)/(2*sd.x^2))))
y <- y - y* (runif(500,min=0.0,max=0.9))

library(quantreg)
library(splines)

model.1.95 <- rq(y ~ bs(x,degree=3,knots=med.x),tau=0.95)
model.1.99 <- rq(y ~ bs(x,degree=3,knots=med.x),tau=0.99)

jpeg(filename="example.cubic.bsplines.jpeg",width=6.5,height=6,units="in",res
=600)
par(cex.lab=1.5,cex.axis=1.5,cex.main=1.5,mar = c(5.1, 5.1, 2.1, 2.1))

plot(x,y,ylab="Biomass",xlab="Mud content (%)")
graph.x <- seq(from = 1, to = 99,by =0.1)
lines(graph.x,predict(model.1.95,newdata=data.frame(x = graph.x)))
lines(graph.x,predict(model.1.99,newdata=data.frame(x = graph.x)))

dev.off()
```


Biomass
Mud content (%)