Reading 4

ES 207

Brendan Smith

**Discussion Questions:**

1. Which work was easier to follow, most relevant to your current research, or interesting? Why?

De’ath and Katharina

1. Where else have you encountered the use of classification/regression trees for data analysis? Have you found them particularly useful?
2. What are some of the limitations you’ve found in the reading or personal experience regarding the use of classification/regression trees?

Cade and Barry

1. What are some of the limitations you’ve found in the reading or personal experience when using quantile regression?

**Summaries:**

"Classification and regression trees: a powerful yet simple technique for ecological data analysis."

Author Backgrounds:

* De’ath – Principal Research Scientist in Statistics, Ecology, Environmental Sciences, Australian Institute of Marine Science. Development of statistical and mathematical models and develops opensource statistical software. Interests include climate change, water quality and starfish in Great Barrier Reef. (http://www.nerptropical.edu.au/people/glenn-death)

|  |  |
| --- | --- |
| h-index | 58 |
| Citations | 9050 |

* Fabricius – Senior Principal Research Scientist, Australian Institute of Marine Science. Researches disturbances (ocean acidification, climate change and terrestrial run-off) in coral reefs both long and short term. Long term exposure of CO2 on coral reef communities, etc. (http://www.nerptropical.edu.au/people/katharina-fabricius)

|  |  |
| --- | --- |
| h-index | 52 |
| Citations | 11067 |

This work, authored by De’ath and Fabricius, is written in the form of a generalized “cook-book” for classification and regression trees. They aim to convince the reader that classification and regression trees are suitable for analyzing complex ecological data that can be nonlinear, high-order and missing values. They performed classification and regression tree analysis on data taken from the GBR for three soft coral taxa.

Classification and regression trees are used to “explain variation of a single response variable by one or more explanatory variables” And we can construct the tree by repeatedly splitting the data into two subets that are homogeneous relative to a response variable at the split. The graphical nature of the trees render them useful for analyzing large data sets. They discuss the exploration and prediction of data using trees, criterion for splitting, simplification by means of cross-validation, and how the tree model can be used to account for transformations and missing values.

In this study, the explanatory variables are cross-shelf position, location and depth. In terms of the classification tree, these can be seen at each split (Fig 2). Mindful that the order of which the splits occur matter. The end result should be a minimum number of splits with much statistical detail. The total sum of squares is represented by the length of each vertical line from one split to the next.

There are many ways to “prune” these trees, or remove splits that are too weak or do not offer much insight to the analysis of the data. One is to allow the tree to grow to a large size, and manually remove these unfavorable splits. Another is to cross-validate these splits computationally, which entails comparing parallel models iteratively, penalizing the model for complexity, or estimating the prediction error. Another way of cross validation is to split the data into several mutually exclusive subsets and build trees based on the same model, then select the tree with the smallest estimated error rate. This is called the V-fold cross-validation.

As stated by the authors, the power of the classification and regression trees are:

1. Wide range of response variable choice
2. Interactive exploration, description and prediction
3. Transformations of explanatory variables do not adversely affect the utilization of classification/regression trees
4. Graphical interpretation of data
5. Ability for cross-validation
6. Ability to efficiently handle missing values without removing data or adversely affecting the remaining data

"A gentle introduction to quantile regression for ecologists."

Author Backgrounds:

* Cade – Biology Research Statistician USGS. Some of his research includes: agriculture, fracking, hydroecology, etc. Current research: Contraception in feral horses, limiting habitat relationships with regression quantiles, quantitative and statistical research, and NWR conservation plan development. No information on google regarding h-index or citation count.
* Noon – Environmental Science. Professor in the Dept of fish, Wildlife, and Conservation Biology at Colorado State University. Full Bright US scholars for the year 2010-2011. Effects of land management on wildlife, mainly imperiled species. Served as Chief Scientist for the National Biological Service, Dept. of the Interior during Clinton Administration. Again, no information on google regarding h-index or citation count.

This work serves as another use-case example of quantile regression in the field of ecology. Quantile regression is a way to glean more information from data sets at each portion of the probability distribution. As the name reveals, the regressions are taken at certain quantiles in the probability distribution. On the first page, an enlightening “In a nutshell” blurb is provided. Some key takeaways from quantile regression is useful for data that has unequal variation due to both measured and unmeasured data, nonlinear relationships such as having multiple rates of change throughout a response, and limiting factors.

The authors first introduce a data set in which they analyze the abundance of Lahontan cutthroat trout to stream width:depth. The relationship was determined to be a nonlinear negative relationship when using the quantile regression method; however, other standard methods would result in an inconclusive relationship between the response and predictor variables.

The initial development of quantile regression was to extend the linear model for estimating the slopes in all parts of the distribution for semiparamtric distributions. They are best developed for linear models, but have breadth to analyze parametric nonlinear, nonparametric, and nonlinear smoothers. But there is no free lunch, there are concerns regarding estimation, inference and interpretation. Thus, making quantile regression a useful tool in the toolbox, but still as bias and error prone as any other regression analysis.

Another favorable attribute of regression quantiles is the ability to nonlinearly transform the data without loss of generality. This was proven useful for another case study in which the survival of Chihuahuan desert annuals were studied. The response variable here is final plant maturity density with a prediction variable of initial germination density.

Utilization of regression quantile entails breaking the interval [0,1] into several unequal length intervals. The number and length of each subinterval is highly dependent on sample size, number of parameters and the measured distribution of the response variable. Regressions are calculated at each subinterval. Sampling variation differs across quantiles, thus is why regression quantiles are a chosen analysis method to begin with; however, as estimates drift further from the median or 50th percentile, the precision decreases. This is shown in the paper (Fig 4) via a confidence band, which can be ween to constrict towards the median, and diverge towards the .90 regression quantile.

One of the powerful attributes of regression quantiles is the ability to detect effects associated with variables that would otherwise have been removed from the data based on relations to the mean. This is especially true in linear models with unequal variances and also large variations, which is common between ecological variables and assumed causal factors. Additionally, as stated previously this tool is extremely valuable when multiple limiting factor interact and not all factors are measured, which is the case in a majority of ecological studies.

**References:**

Cade, Brian S., and Barry R. Noon. "A gentle introduction to quantile regression for ecologists." *Frontiers in Ecology and the Environment* 1.8 (2003): 412-420.

De'ath, Glenn, and Katharina E. Fabricius. "Classification and regression trees: a powerful yet simple technique for ecological data analysis." *Ecology*81.11 (2000): 3178-3192.