Simple Linear & Non-Linear Regression

What Is It?

N-STEPS Objectives

- Provide regions, states, and tribes with support related to nutrient criteria development
- Provide access to expert assistance with issues related to nutrient criteria development and implementation
- Improve communication nationwide.

One of the most common statistical modeling tools used, regression is a technique that treats one variable as a function of another. The result of a regression analysis is an equation that can be used to predict a response from the value of a given predictor. It can be used to consider more complex relationships than correlation by using more than two variables or combinations of different order equations (e.g., polynomials). Regression is often used in experimental tests where a range of fixed predictor levels are set and one tests whether there is a significant increase or decrease in the response variable along the gradient of predictor levels.

Example Types (*treated in this fact sheet):

- Simple linear*
- Multiple linear regression
- Non-linear
  - Logistic regression*
  - Exponential regression
  - Polynomial regression

Example Question: Can I use total phosphorus concentration to determine the chlorophyll content in a lake?

How is it Applied to Nutrient Criteria Development?

Nutrient criteria development involves three main processes: identifying relationships between biological responses and nutrient stressors, examining these relationships, and establishing nutrient and/or biological thresholds or criteria.

If a strong relationship between a biological parameter (e.g., algal biomass) and nutrient variables (e.g., total phosphorus) is or is not identified in a correlation analysis, scatterplots and regression analysis can be used to examine the relationship further. Regression analysis includes simple linear regressions, multiple linear regressions, and non-linear regressions. Simple regression analysis is similar to correlation analysis but it assumes that nutrient parameters cause changes to biological attributes. Nonlinear or multiple linear regression analyses can be used to consider more complex relationships between biological attributes and nutrient variables, such as nonlinear relationships and multiple predictors (e.g., both TN and TP are predictors of algal biomass).

How Does It Work?

Simple linear regression – In least squares regression, the common estimation method, an equation of the form: $E(y_i) = \beta_0 + \beta_1 x_i$ is estimated by finding values for the parameters (β_0- the intercept and β_1- the slope) that minimize the sum of the squared deviations between the observed responses and the linear equation. The variance of each parameter can be used to evaluate its significance. A significant slope means the slope is different from zero and there is a response to the predictor; a significant intercept means the intercept is generally different from zero.

Some Assumptions:

- Relationship between predictor and response is linear (n.b.: transformations used to make it linear)
- Error term is assumed to be normal (bell shaped distribution) with homogeneous variance -
- Samples are independent

Logistic Regression – Logistic regression is used to model a binary response (e.g., presence/absence of nuisance algae) with some predictor (e.g., nutrient concentration) using an equation of the form:

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

. The probability of response is modeled rather than the actual response value, typically using a computationally intensive iterative process. Logit ($\log_e$ of odds ratio) gives a linear model.

Some Assumptions:

- Error term is assumed to be non-normal with nonconstant variance
- Samples are independent

Data Requirements

Independently collected numeric data in the form of paired observations are required – for both the predictor(s) and the response variable. These are typically numeric data, although discrete numeric and binary variables (presence/absence) can also be used. As with correlation, the greater the range of environmental conditions encompassed the better. One way to assure a large range is to use a gradient design and select sites along as large a gradient as possible.

What Should You Look For & Report?

Linear Regression – Examine the plots and the final regression line. Examine the residuals of the regression for normality (equally spaced around zero), constant variance (no pattern to the residuals), and outliers. Report the regression equation, the significance of the model, the degrees of freedom, and the significance of each of the parameters (t-statistics and p-values for the slope and intercept). It is not uncommon to also report any unusual features of the residuals, if they exist. Finally, report the coefficient of determination (R^2) which measures what proportion of the variability in the relationship is explained by the regression. This varies from 0 to 1, where 1 means the regression explains 100% of the variability in the relationship (i.e., all the points fall right on the regression line).

Logistic Regression – Examine the plots and final regression line. Use a goodness-of-fit test to determine the appropriateness of the model. Fitted responses should approximate monotonic curves with sigmoidal shapes. Formal and informal tests of this are available (e.g., Hosmer-Lemeshow). Report the parameter estimates and the associated significance, and the goodness-of-fit results. One can use the parameters to determine the probability of a response for a particular value of the predictor.

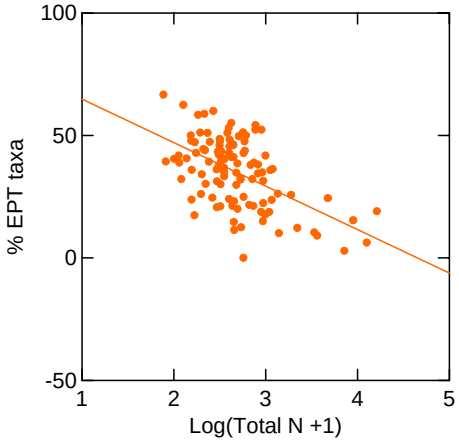
For both forms, avoid extrapolation – predictions beyond the range of the predictors used to build the models, as confidence outside that range is low.

Pros

- Efficient analysis that is useful for prediction
- Easy to interpret
- Can use more than one predictor
- Synergistic relationships can be modeled

Cons

- Assumptions about variables constrain analysis
- Evaluating models becomes more difficulty with complexity of model
- Shape of response necessary to choose the best model
- Sensitive to outliers and, like most models, hard to extrapolate



% EPT = 82.6 – 17.8*Log(TN)
F = 44.2, p < 0.001
df = 106
R^2=0.29

Alternatives

Generalized Linear Models
Generalized Additive Models
Nonparametric regression

Citations

EPA Statistical Primer - http://www.epa.gov/bioindicators/statprimer/index.html

Ott, R.L. 1993. An introduction to statistical methods and data analysis. 4th edition. Duxbury Press, Belmont, CA.