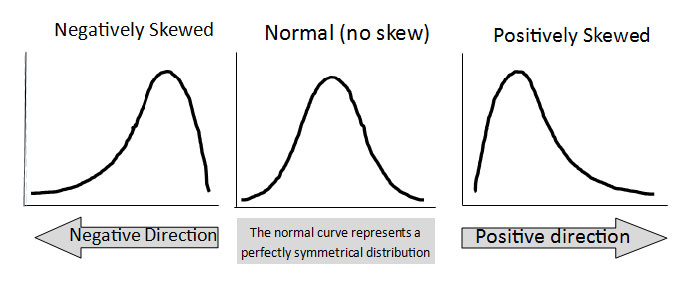
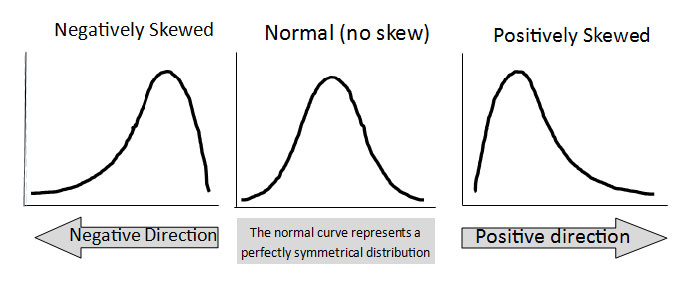
Basic Practices

* Projects should be easy to read – ANNOTATE WELL for good source code
  + Use comments often (be consistent)
    - # end of line comments
    - ### major sections – creates modules
  + Use “<- “more often than “=”
* Learn about the package you are using and how to cite (Vignettes)
  + “Cran \*insert package name\*”
* Use source code (helps explain what is happening “under the hood”)
  + Type the function name in R console to return source code
* Good practices
  + Work with source files
  + Keep source code “readable”
  + Read documentation
  + Learn from the best – read R package sources
  + Do not copy and paste (ex. use “for loop”) – create small modules (###)
  + Strive for clarity and simplicity
  + Test code!
* Naming Code
  + Be able to search for files easily – using default ordering
    - Numeric First
  + Year-Month-Day (2024-01-16): ISO 8601 standard
  + Case sensitivity is important
  + Use “-“ to separate words
  + Use “\_” to separate sections

Linear Regression Review + Data Transformation

* Assumptions of Linear Models (LINE)
  + L: Linearity – the *means* of the subpopulations of Y all lie on the same straight line
  + I: Independence – the Y values are statistically independent
    - Ex. same species data might *not* be independent or individuals
  + N: Normality – subpopulations of Y values are normally distributed (bell curve), for all values of y at x
  + E: Equal variances – variances of subpopulations of Y are all equal (homoscedasticity)
    - Ex. bell curve at every value of x but the variances cannot vary
    - Heteroscedasticity can be dealt with using log(Y) or log-log
  + If data violates assumptions:
    - Option 1: use a more appropriate/complicated model
    - Options 2: (ideal) find a transformation of data (ex. log, x2, x1/2)
* Terms to know:
  + Transformation – application of same calculation to every point of the data separately
    - Non-linearity dealt with by transforming X values (ex. x2, SQRT)
  + Normalization – process of scaling so data has range from 0 to 1, necessary in multiple linear regression (ex. 5 predictors to be on the same scale to interpret easier)
  + Standardization – transforming so data has mean of zero and SD of 1
  + Residual – difference between the predicted values and the actual values of response var.
* How to Choose Transformation:
  + Check for assumption violations
  + Understand/accept that your new scale of measurement for data is allowed
  + Use Q-Q Plot (sample quantiles vs. theoretical quantiles) with Gaussian Distribution
  + Evaluate via model diagnostics, *not model results*
  + Types of transformations:

|  |  |
| --- | --- |
| **If your data distribution is:** | **Use this transformation method:** |
| Moderately positive skewness | X = SQRT(X) |
| Substantially positive skewness | X = log10(X) |
| Substantially positive skewness (w/zeros) | X = log10(X + C) |
| Moderately negative skewness | X = SQRT(K - X) |
| Substantially negative skewness | X = log10(K - X) |

\*C & K – constant so smallest score is 1

A diagram of a function

Description automatically generated

Generalized Linear Model

* GLM = important for data that cannot be easily transformed to meet normality
* Assume: cause + effect relationship between X & Y,
* What is actually happening? We are using a ‘link’ function with parametric statistics, and we re-specify the error distribution
* 3 Components
  + 1. Random(variance) = response (Y), normal distribution with error identically distributed
  + 2. Systematic = predictor (X), linear in parameters
  + 3. Link = “identity” of Y or normal value without transformation, specifies link between , modeling mean directly
* 3 Steps
  + 1. Specify distribution of dependent variable (binomial, poisson, etc.)
  + 2. Specify link function (logit, log, etc.)
  + 3. Specify linear predictor relates to independent variables (slope and intercept)

\*insert image\*

* Link functions
  + Logit = X is categorical
  + Probit = X is continuous
* Poisson Regression
  + Count data (0-10 ish); discrete
  + Lambda
  + 0 inflated data = observation data
* Use R help page “?family” for more information (especially about default settings)
* Advantages
  + Do not need to transform the data (impossible for binary data)
  + Choice of link function can be independent from random component – but usually don’t
  + Keep original combinations of linear predictors
  + Models are fitted with Maximum likelihood estimation
  + One procedure to model using whole family of link functions (automatic)
* There is a difference between general and generalized
  + General: linear regression, ANOVA, analysis of covariance, multiple linear regression
  + Generalized: logistic regression, binomial regression, log-linear models
  + If you expect it to be normal/gaussian and it is not (non-linearity), use generalized model

Ordinations & Multivariate Statistics – S. Koerner

* 3 Categories of analysis
  + Univariate (1 variable = petal length), bivariate (2 variables = x vs. y), multivariate (3+)
* Multivariate statistics: broad techniques, descriptive and inferential, confusing terminology
  + Purpose: describe a community in the field with multiple variables
* Ordinations used to visualize patterns seen with multivariate statistics
* Common types: Species by site matrix, site environmental data with niches, site spatial gradients, remove rare species from data (controversial)
* Ask: Why use ordination? What variables you have?
  + Determine use of constrained or unconstrained ordination tests (complementary)
* Types
  + Unconstrained = maximize explained variation in community data, “where do we get the most spread in the data” – why is each point different from the other
    - PCA: Principle Component Analysis = analysis of variance and creates a smaller set of components with the same variances (x & y independent of each other, unrelated), axis 1 and 2 show the most variance so you don’t retain all axes created, best for *environmental* data where you need to reduce the number of variables, need more sites than variables
    - NMDS: Non-metric multidimensional scaling = for ecological different species data within a site(s),
      * construct (dis)similarity matrix – calculates how far apart each site is (site by site matrix)
      * taking multidimensional plot and makes a 2-D plot with largest spread between variables
      * rank-based approach, iterative algorithm, no unique solution, stress values
    - Post-hoc tests on dissimilarity matrix
      * Anosim – analysis of similarity
      * Adonis – examining group means and variance
      * Betadisper – how much variance between/within plots
  + Constrained = maximize explained variation
    - Linear response model
    - Unimodal response model
* R package = Vegan: Analysis for Ecology Community Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique​** | **Objective​** | **Dependence Type​** | **Data Structure​** |
| Unconstrained Ordination (PCA, PO, CA, DCA, NMDA)​ | Extract gradients of maximum variation​ | Interdependence​ | One Set; >>2 variables​ |
| Cluster Analysis​ | Establish groups of similar entities​ | Interdependence​ | One Set; >>2 variables​ |
| Discrimination (MRPP, MANTEL, DA, CART, ANOSIM, PERMANOVA)​ | TEST for or describe differences among groups of entities or predict group membership​ | Dependence​ | Two sets; 1 grouping variable, >> 2 discriminating variables​ |
| Constrained Ordination (RDA, CCA, CAPS, CanCorr)​ | Extract gradients of variation in dependent variables explainable by independent variables​ | Dependence​ | Two sets; >>2 dependent variables, >> 2 independent variables​ |

New Title



New Title