Data Workflow

## TIPS

1. Record your data workflow whether it is in an Excel document or within your code. If you do something not in your code file, document it somewhere.
2. Organize your analysis project folder well.
3. It is also a good idea to keep track of your filenames and a summary of what they do so you can find your relevant analyses when you are looking for them.

## Data Types

Environmental Data (continuous and discrete)

Biological Data (multiple species, multiple programs)

## Data Exploration

1. Decide on variables of interest
2. Are there outliers that need to be cleaned up?

* Use known ranges for variables (e.g. water quality)
* Cleveland dotplots
* Boxplots
* Histograms
* Plotly for exploration

1. Are there correlations between environmental variables?

* Correlation matrix or pairplots
* Corvif > 4.0 = multicollinearity

1. Is there a relationship between your response and predictor variables?
   1. Correlation matrix or pairplots

## Data Cleanup

1. Missing data

* Impute mean based on factor that makes sense
* Model the missing data based on variable that makes sense

1. Outliers

* Decide if outliers are “real” – in general, should not delete it if it is likely to be a real observation
* Edit data, flag, or remove from analysis
* Flagging generally good for publishing data, but remove for analysis

## Types of analysis

### Monte Carlo

Data are randomized and reshuffled (simulated dataset). Null hypothesis states that this random distribution is not different from the pattern in the actual data.

**Assumptions:**

1. Independence: the data collected represent random, independent samples
2. The test statistic describes the pattern of interest
3. The randomization creates an appropriate null distribution for the questions

*Don’t know much about this one!*

### Parametric

**Assumptions:**

1. Normality
   1. Not too important unless very non-normal. Violation okay if sample size is large enough
   2. Pooled residuals should be normal
2. Homogeneity: spread of data should be the same at each X value
   1. Minor violation okay
   2. Pool residuals, plot against fitted values
   3. Bartlett’s test (sensitive to non-normality, but useful for testing effect of transformations)
3. Fixed X (explanatory variables): explanatory variables should be known in advance (experiment)
   1. Violation is okay if the error around your explanatory variable should be small compared to the range of your explanatory variable
4. Independence: the Y value at Xi is influenced by other Xi (observations are close in space or time)
   1. Nature of the data:
      1. E.g. if there are salmon present today, it is more likely there were salmon present an hour ago. If there are salmon at one spot, it is likely there are salmon just upstream.
      2. Solution: add temporal or spatial dependence structure between observations (or residuals) in model
   2. Improper model:
      1. Plot residuals vs. X – if there is a pattern, then violation)
      2. Need to improve model or transform to linearize relationship
5. Correct model specification

### Non-Parametric

Based on analysis of ranked data.

**Assumptions**:

1. Independence: independent and random sampling
2. Data in different groups have the same distribution as each other (skew left vs right, kurtosis),

Mann Whitney U (two groups of continuous data, like t-test) – for independent samples   
Wilcoxon signed rank test (paired t-test) – for repeated measures on the same subjects   
Kruskal-Wallis (three groups of continuous data, like one-way ANOVA)  
Spearman rank correlation (linear regression/ correlation)

Mann-Kendall: detect monotonic trends in series of environmental data, climate, data, or hydrological data.

* Does a time series have a monotonic (non-changing) upward or downward trend?

### Bayesian

Use prior information/distribution to improve analysis, Determine the probability of the hypothesis given the data that have been collected.

Really don’t know much about this one!

## analysis

### Parametric

#### linear regression

* Continuous variables, Y = f(x)
* Fixed term describes response variable as a function of explanatory variables
* Random term contains components allowing for:
  + heterogeneity, (generalized least squares)
  + nested data (random effects), (mixed effects)
  + temporal correlation, spatial correlation,
  + real random term (linear model)

#### T-Test

**Format**: Predictor: categorical, Response: continuous – 2 treatment groups   
**Types**

* One-sample: Is the sample mean statistically different from the population mean?
* Independent: Effect of one variable on two populations
* Paired: Effect of treatment (pre-post) on a single population

#### ANOVA

**Format:** Predictor: categorical, Response: continuous – 3+ treatment groups  
**Example:** Effect of site on CPUE at three locations/ Effects of site and season on CPUE

**Types**

* One-way (one predictor), Two-way (two predictors)
* Split-plot: one factor applies to each group, another factor applies to all groups
* Repeated measures: multiple observations are taken on a single individual or replicate
* Fixed factor:
* Random factor:

**Post-Hoc**<https://stat.ethz.ch/~meier/teaching/anova/contrasts-and-multiple-testing.html#bonferroni-holm>

* T-tests, need to add a correction to the p-value:
  + Bonferroni: experiment-wide error rate/number of tests
  + Dunn-Sidak:

#### ANCOVA

* Predictor: categorical and continuous, Response: continuous

#### Kolmogorov-Smirnov Test

**Format:** Continuous data

**Example:** Compare the size distribution of male to female in a species

* **One sample:** Compare the distribution of a variable to a known distribution
* **Two sample:** Compare the distribution of two samples to each other

### Categorical data

#### Contingency tables

Chi-Square

## Model validation

### Parametric/non-parametric

**Model Validation:**

1. Residuals vs. fitted values to verify homogeneity
2. QQ plot of histogram of residuals to verify normality
3. Residuals vs. each explanatory variable to check independence
4. Check for influential observations (leverage)

#### If you run into violations

##### Try:

* Add Interaction terms
* Extend model with non-linear effects (e.g. length ^2 rather than length)
* Add more explanatory variables
* Transform data
* Try GAM

##### Transformations and Standardizations

1. Transformations
   1. **May help meet the assumptions of normality and homogeneity**
   2. Highly skewed variables
   3. Better meet assumptions of statistical test
   4. Emphasize presence/absence
   5. Equalize relative importance of variabilities
   6. Do not change their rank – use power, log, arcsine, arcsine sqrt
      1. Log or sqrt for highly skewed or ranging >2 orders magnitude
         1. Log: good if mean and variance are positively correlated, data that are skewed right
         2. Sqrt: often for count data. Sqrt(x + 0.5) so that 0 data gets transformed as well.
      2. Arcsine square root for proportion, percentage data
      3. Reciprocal transformation
         1. Used for data that record rates (e.g. number of offspring per female): 1/X.
      4. Consider binary transformation when percent of zero values is high (>50%) or distinct values low (<10)
   7. Use same transformation for same variable set (e.g. species)
   8. Remember to back-transform data and confidence intervals
2. Standardizations
   1. Place highly unequal sample units or variables on equal footing
   2. Better represent patterns of interest
   3. Type depends on objective and what makes sense
      1. Z-score
         1. different scales or units. Highest weight to those with highest variability.
         2. (Value – mean column) / sd column
      2. Total
         1. Equalizes heights of peaks of species response curves, good for extreme values; Can emphasize rare species
         2. Value / max(column)
   4. Column standardization
      1. “equalize” variables measured in different units and scales
   5. Row standardization
      1. Species datasets (e.g. relative abundance within sites)

##### For violations of heterogeneity

* allow for different variances in the model (generalized least squares estimation)
* use a different distribution and model structure (Poisson, negative binomial, Gamma distributions GLM)

##### For violations of independence

* Incorporate temporal or spatial dependence structure between observations (or residuals) in model)
* Smoothing methods

##### Other

* Consider nested data and random effects

## Model selection

AIC

## Multivariate Analyses

Unconstrained vs. Constrained Ordination:

* **Unconstrained**: Variability in one set of data (no predictor vs. response) – describes variability in data, but not effect of one group of variables on another
  + - * PCA, PCoA, NMDS
* **Constrained**:Effect of one set of multiple variables (predictor) on another (response)
  + - * CCA, RDA

### Principal Components Analysis (unconstrained – one dataset)

**Format:** One dataset of variables (no predictor/response) – continuous, count, categorical

**Example:** Environmental descriptors of a dataset, Morphometric/Systematic data. Not good for biological data with lots of zeros.

**Why:** Reduce dimensionality of your dataset by placing similar objects in closer to each other in ordination space. Condenses large number of variables into smaller set with minimal loss of information.

* Fewer correlated variables – condenses all the information
* Shows which variables are primary drivers of variability

**Assumptions:**

1. Multivariate Normality
   1. Univariate normality of original variables
   2. Normality of PC scores
2. Few outliers
   1. Do some outlier detection diagnostics
3. Linearity: variables change linearly along underlying gradients, linear relationships exist among variables
   1. Check for horseshoe effect (result if not linear)
4. Sample size: More samples (rows) than variables (columns)
   1. N>3\*P

**Methods:**

* Decide if correlation or covariance matrix
  + Correlation: variables are different units or scales
  + Covariance: variables are the same unit
* Determine number of Principal Component Axes you want to use (PCs)
  + Latent Root Criterion: eigenvalue > 1 for correlation matrix
  + Scree Plot: use shape of curve (stop when levels out)
  + Broken Stick: Eigenvalues > broken stick distribution are kept
  + **Randomization**: Randomly shuffle objects n number of times (e.g. 1000). Compare each PC eigenvalue to distribution of eigenvalues under null hypothesis (no correlation).

**Interpretation:**

* Plot PC scores in first few dimensions.
  + Points similar to each other will be closer to each other.
  + Points 180 degrees from each other are negatively correlated, whereas 90 degrees away have no correlation
  + Can color-code by variable to see which samples are driven by which environmental variables
* **Which variables are driving the new PC axes?** 
  + Loadings are bivariate correlations between variables and the PC. Higher loadings indicate higher correlations.
  + Loading > abs(0.30) are considered significant need to check
* **How well does the PCA do?** 
  + Shepard Diagram: Compute distance between all pairs of objects – correlate and plot distance in reduced space vs. distance in multivariate space

### NMDS

**Format:** One dataset of multiple variables (e.g. species)

**How it works:**

* Iterative refining of positions of objects in multidimensional space until they are close to ranked dissimilarity relations between objects
* Attempts to minimize **stress** – correspondence between ranked distances and ranked multivariate distances

**Pros:**

* No underlying assumption of linearity
* No assumption of linear or unimodal response function
* Can use any dissimilarity measure
* Appealing for visualizing patterns of relative differences (though these are **not** absolute differences)

**Methods**:

* Transform data if deemed useful (e.g. species counts can be skewed)
* Conduct NMDS
* If stress is high, try increasing k (number of dimensions)
* You can use a scree plot to determine number of dimensions you want to use

**Interpretation:**

* **How well did the test perform?**
  + Use Monte Carlo randomization test on final stress value
  + Stressplot: check correlation between calculated dissimilarities and plotted values
* **Visualize results**
  + Plot NMDS – this will display all your individual samples
  + To examine which species correspond with which axis, use envfit: plot species loadings

### PERMANOVA (Non-parametric, based on distance matrix)

**Format:** Effect of one+ factors on several response variables

* Matrix of continuous variables and predictor variables (species + factors are column headers)

**Example:** Effect of season and site on suite of species

**Why:** Similar to an ANOVA. What is the effect of certain variables on the combination of variables (e.g. species composition)?

**Assumptions:**

* Overall pretty flexible! No distribution assumptions.
* Observation units are exchangeable under a true null hypothesis.
* Some level of independence needed, especially if variables are correlated.
* Check for homogeneity of variance even though not an explicit assumption, since significant difference between groups could be due to differences in spread or location in multivariate space
  + - * ANOVA: Distance between group members to group centroid

**Methods:**

1. Standardize species data to relative species abundance: Rescale each row so that sums equal 1
2. Conduct PERMANOVA. Pick dissimilarity matrix (usually bray for ecological data)
   1. If multiple variables and you don’t have specific hypothesis, run PERMANOVA for each individual variable, then pick the one with highest R2 that is significant and add on variable with next highest R2. Continue this process until only significant variables are added.
3. Check for multivariate homogeneity of group dispersions (variance)
   1. Anova
   2. Permutational test
   3. Plot dispersions

### Canonical Correspondence Analysis

Multiple species biological observations ~ Multiple environmental parameters

### RDA

## Generalized Linear Models

**Difference between linear model and mixed models:** Linear mode relies on certain assumptions (e.g. normality, homogeneity of variance, fixed X, independence, correct model assumption). If violated, explore mixed models.

Count or CPUE ~ Multiple environmental parameters  
Length or Weight ~ Multiple environmental parameters

**Overdispersion**: Variance is larger than the mean

* Residual deviance/degrees of freedom
  + Check anova(M1)

### Zero-Inflated Poisson

Good for count data, since lots of zeros.

#### Finding the best model

* Automatic forward, backward selection
  + Step(M1)
* Hypothesis testing
  + Test the null hypothesis using z-statistic, drop least significant term, refit model, keep dropping until all terms significant
  + Use drop1(M1, test = “Chi”), applies **analysis of deviance \* Recommended**
  + Use anova(M1), applying analysis of deviance to each sequential drop

### Zero-Inflated Negative Binomial

Good for count data, since lots of zeros. Allows for greater overdispersion.

## Generalized Additive Models

Count or CPUE ~ Multiple environmental parameters  
Length or Weight ~ Multiple environmental parameters