An Introduction to **R** for Biomathematics

Christopher J. Mecklin
Department of Mathematics & Statistics
Murray State University

May 16, 2008

Outline of Talk

- Introduce the **R** statistical programming environment
- Discuss the advantages and disadvantages of **R** versus other statistical software
- Demonstrate using **R** for standard statistical analyses

Outline of Talk

- Demonstrate the use of **R** in three situations where “standard” statistical analysis is either not possible or not advised
- Briefly discuss other possibilities with **R**

What is **R**?

- **R** is an open-source statistical programming environment that is available for free.
- **R** is similar to S, a statistical programming language originally developed at Bell Labs.
- **R** was originally develop at the Univ. of Auckland (NZ) by **R**obert Gentleman and **R**oss Ihaka and has been maintained by a core group of use**R**s since 1997.
- In addition to the core group, many users have added to **R** by submitting packages to perform many types of statistical tasks.

What is **R**?

- The syntax for **R** is quite similar to C++ and it is an object-oriented language.
- This is in contrast to SAS, whose DATA steps and PROC statements are much more reminiscent of FORTRAN.
- Note: FORTRAN was a programming language that us old people learned in Computer Science 101 in the 20th century while listening to our SONY Walkman and having to wait until we got back to our dorm room before being able to call our girlfriend/boyfriend.

CRAN (The Comprehensive **R** Archive Network)

- http://cran.r-project.org
- Or just type “CRAN” into Google!
- Versions available for Windows, Mac, Linux
- It’s free!!!
- Yes, it’s free!!! Not only is it free, it’s good!!!
- The BioMAPS group will purchase **R** for you.

CRAN (The Comprehensive **R** Archive Network)

- I will load the "**BiodiversityR**" package (I just noticed this a few days ago, as it has only existed on CRAN since 5/9/2008)
- If a package you are downloading requires other package(s) that you have not yet downloaded, **R** will automatically download them for you!
- I've been using another package called "**vegan**" for calculating measures of species diversity
- More on this later

Too Many Packages!!!

- New packages are being developed constantly and it’s hard to keep up with them.
- http://cran.r-roject.org/web/views/
- We’ll take a quick look at the “**Environmetrics**” views page.
- Others might want to look at the “**Genetics**” or “**Spatial**” page.

Journal of Statistical Software

- A special issue of this e-journal was devoted to articles involving the analysis of ecological data using **R**.
- Later we will look a bit at the **ade4** package, developed by French ecologists Stephane Dray and Anne-Beatrice Duffour. This was one of the articles from this special issue.

Advantages of **R**

- Object-oriented language similar to C++
- Good graphics
- Hundreds of packages available for many specialized forms of data analysis
- Would you pay $200, $300, $400 for **R**? You don’t have to, it’s FREE!!!

Disadvantages of **R**

- Object-oriented language similar to C++ (it's hard to get use to if you are used to working with SAS)
- Relatively steep learning curve
- Some packages have better documentation than others

Disadvantages of **R**

- Some duplication of procedures from package to package and potential for confusion.
- The graphics can be hard to make, save, and export to another document.
- Creating and importing data sets can be a real pain!
- By the way, I could make most of these complaints about SAS as well.

Alternatives to **R**

- If your data analysis needs are modest (STAT 101 level), you might be happy with a calculator or spreadsheet. You probably already have this stuff.
- If your data analysis needs are more substantial, commercial packages like SAS, S-Plus, SPSS, Minitab, etc. are available. Generally expensive.
- If your data analysis needs are fairly specialized, a specialty program like MARK, PAST, CANOCO, ADE, RT etc. might be called for. These packages range from free to expensive and from simple to excruciating in ease of use.
- Many of these esoteric programs are being replaced by **R** packages.

Using R for standard analysis (multiple regression)

- We will look at a quick "textbook" example of a standard multiple regression Descriptive stats, boxplot, histogram
- MLR, corr matrix
- Residuals, diagnostics
- IF you like doing the matrix algebra "yourself" you can do this with **R** similar to SAS/IML (I find SAS/IML to be infuriating!)
- If you would like to "convert" to **R**, check out the Quick-R homepage, http://www.statmethods.net

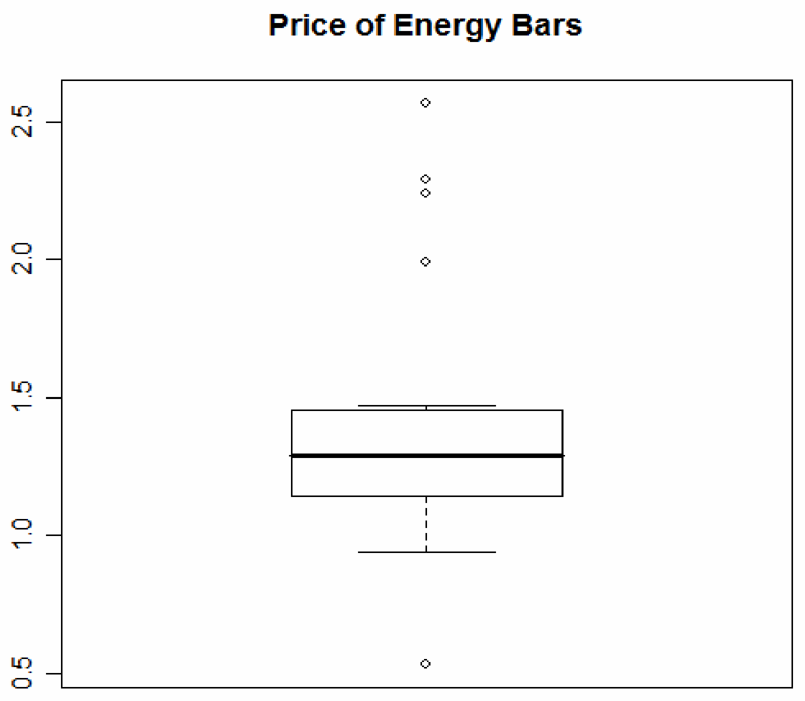
Energy Bar data set

- The following “textbook” data set considers the price of energy bars as the response variable Y, with the calories, protein, and fat content as potential predictor variables X.
- I will quickly demonstrate some standard statistical options but will not attempt a full proper analysis of this data (i.e. I’m going to ignore outliers and possible violations of regression assumptions)

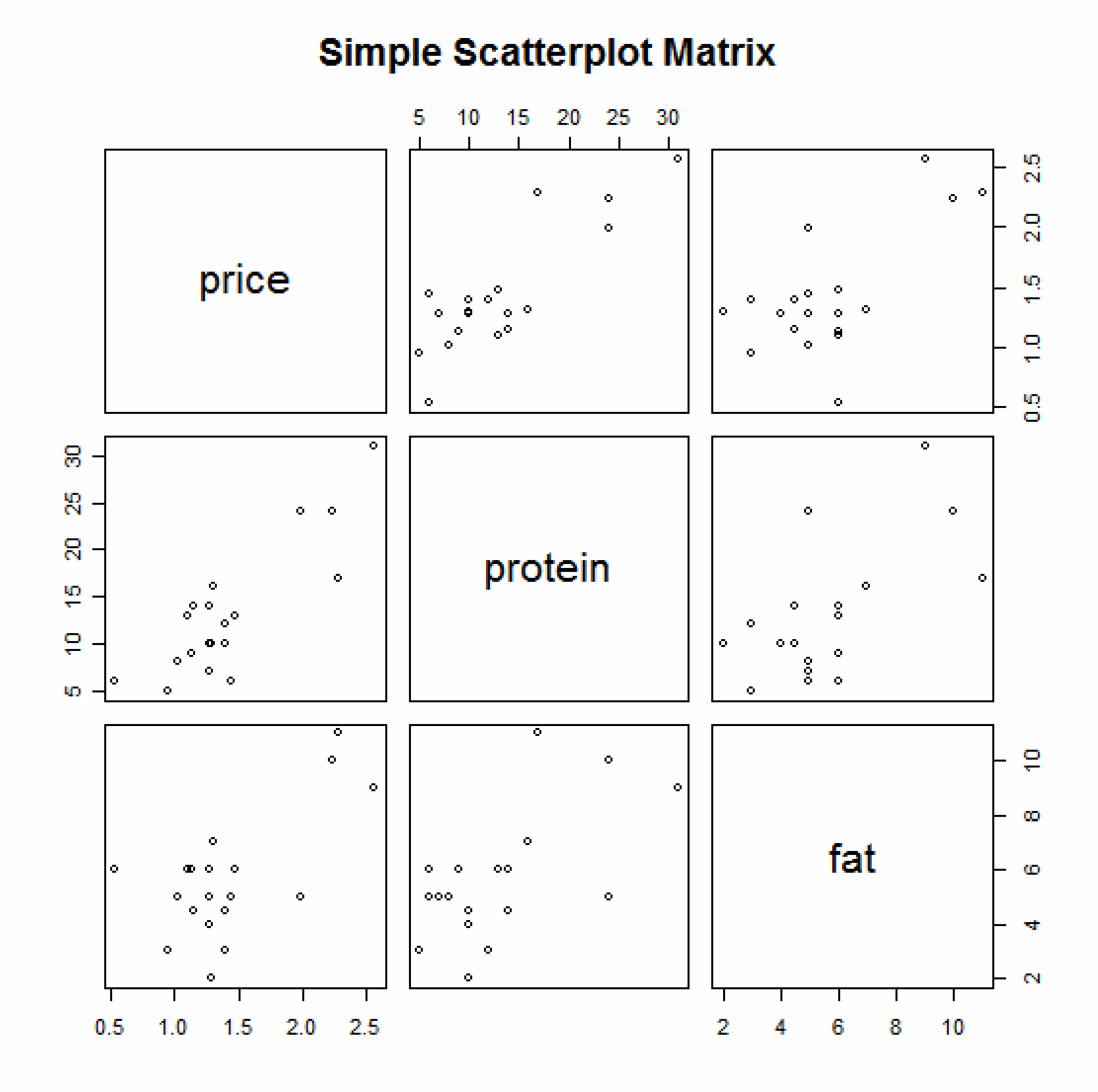
Descriptive Statistics

- R>summary(price)
- Min. 1st Qu. Median Mean 3rd Qu. Max.
- 0.530 1.140 1.290 1.426 1.455 2.570
- R> sd(price)
- [1] 0.5061973
- R> cor(price,calories)
- [1] 0.4954693

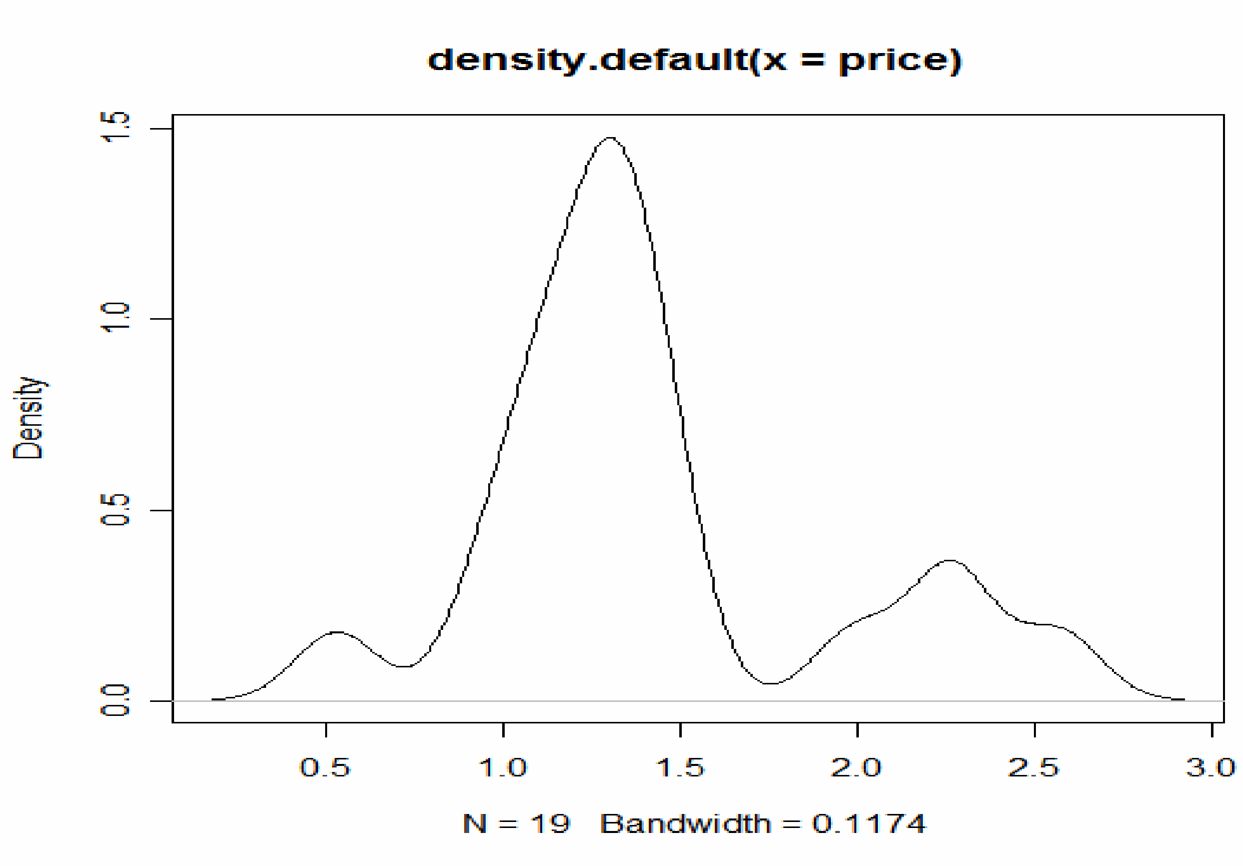
Boxplot



Correlation Matrix



Density Plot of Price



Regression Results

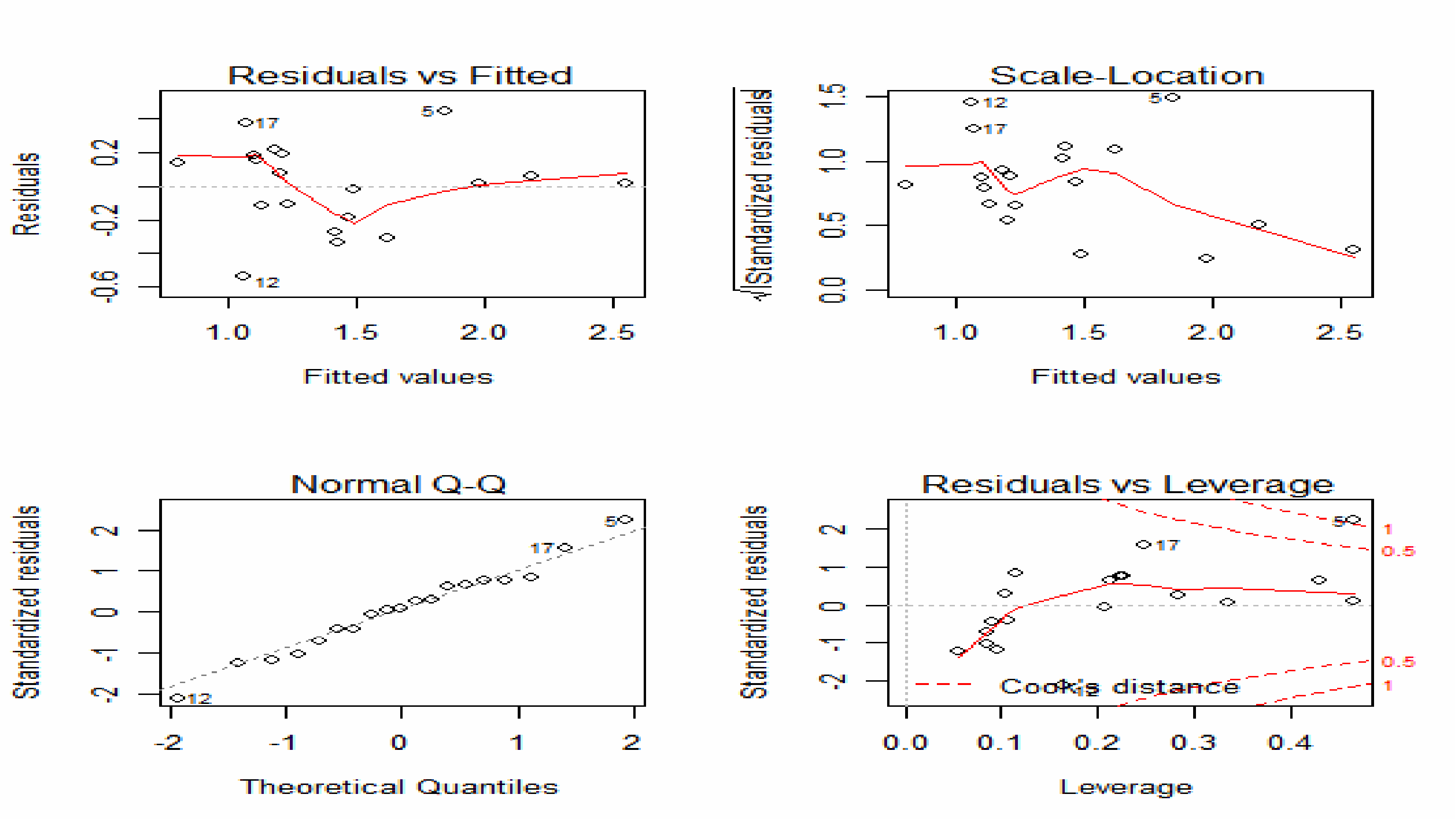
```
Call:
lm(formula = price ~ calories + protein + fat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53343 -0.14946  0.01815  0.16864  0.44674

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3253830  0.3479284   0.935  0.36450
calories    0.0008163  0.0016584   0.492  0.62970
protein     0.0501396  0.0132444   3.786  0.00179 **
fat         0.0456595  0.0356516   1.281  0.21974
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2745 on 15 degrees of freedom
Multiple R-Squared: 0.7549,     Adjusted R-squared: 0.7058
F-statistic:  15.4 on 3 and 15 DF,  p-value: 7.559e-05
```

Some Regression Plots



R vs SAS

- SAS loves to give you reams of output even and especially when you don't want everything. You have to ask it (with noprint options) to suppress output.
- **R**, on the other hand, tends to be much more sparse with results. You have to specifically ask for stuff (like the predicted values, ANOVA table, etc.)
- Which approach is better? Potato, potato

What if standard analysis is not possible?

- There are many, many types of analyses that simply cannot be performed with basic tools such as graphing calculators or spreadsheets.
- There are even many, many types of analyses that still cannot be preformed (without some time-consuming programming by the user) on commerical packages such as SAS, SPSS, Minitab, etc.

What if standard analysis is not correct?

- Another problem is that many users of statistical methods will always try to use methods that they know (regression models, ANOVA, etc.) even in situations where those methods are not the best choice or even appropriate.
- A common theme of my consulting work with smart people from other fields is when they have a situation where they know a standard regression/ANOVA is not the way to go, but they don’t know exactly what to do.

Three research problems

- Species Diversity with the **vegan** package (also the brand-new **BiodiversityR** package that I just discovered and haven’t used)
- Analysis of Left-Censored Environmental Data with the **NADA** package (i.e. nondetectable elements/compounds)
- Ordination with the **ade4** package

Species Diversity

- Biodiversity has been defined to be ``an average property of a community” (Patel & Taillie, 1982) where the property in question isspecies rareness.
- In a diverse community, the typical species is rare; that is, the relative abundance of that species make up a small proportion of the population).

Species Diversity

- A wide variety of indices exist for the measurement of ecological diversity.
- Common choices include the Shannon index and the Simpson index. Both of these indices are a function of the proportion of individuals found in each species.
- Unfortunately, the arbitrary selection of diversity index can lead to conflicting results.

Shannon’s Measure of Species Diversity

- The Shannon index is probably the most commonly used measure of diversity (Shannon & Weaver, 1949)
- Define the Shannon index as:

$$H' = -\sum p_i \ln p_i$$

Simpson's Measure of Species Diversity

- The Simpson index is another common species diversity measure.
- Define the Simpson index as:

$$s' = 1 - \sum p_i^2$$

- There is no reason to necessarily prefer the Shannon index over the Simpson (or vice versa).
- In most situations, the sample with the highest diversity as estimated by the Shannon index will also have the highest diversity when the Simpson index is used.

Renyi’s Generalized Measure of Species Diversity

- Tothmeresz (1995) suggested the use of Renyi's entropy.

$$H_\alpha = \frac{\log \sum p_i^\alpha}{1-\alpha}$$

- Shannon's index is a limiting function of this measure as alpha approaches 1.

Renyi Diversity Profiles

- He suggested plotting diversity profiles (varying the value of alpha) using the Renyi's index family.
- If the curve for a sample lies above the curve of the other sample over the entire range, then the first sample can be said to have higher diversity.
- However, when the curves intersect, particularly if the intersection occurs when alpha is between 1 and 2, then the samples are said to be non-comparable.

Dinosaurs Example

- Fossil data was collected in North Dakota and Montana concerning the diversity of families of dinosaurs for during the Cretaceous period (Sheehan, et al. *Nature* 1991)
- It was concluded that the biodiversity of dinosaurs had remained relatively constant and had not been declining before mass extinction was caused by a cataclysmic event (such as a collision with a giant meteorite).

Criticism of Dinosaur Study

- The original study used standard ANOVA models for analysis, failed to reject the null hypothesis (i.e. equal diversity for each portion of the Cretaceous period), and made the error of concluding this failure to reject was evidence for a constant level of dinosaur biodiversity.
- There have been several re-analyses of this data set with a variety of more sophisticated methods, which did yield the same conclusion.

Dinosaur Data

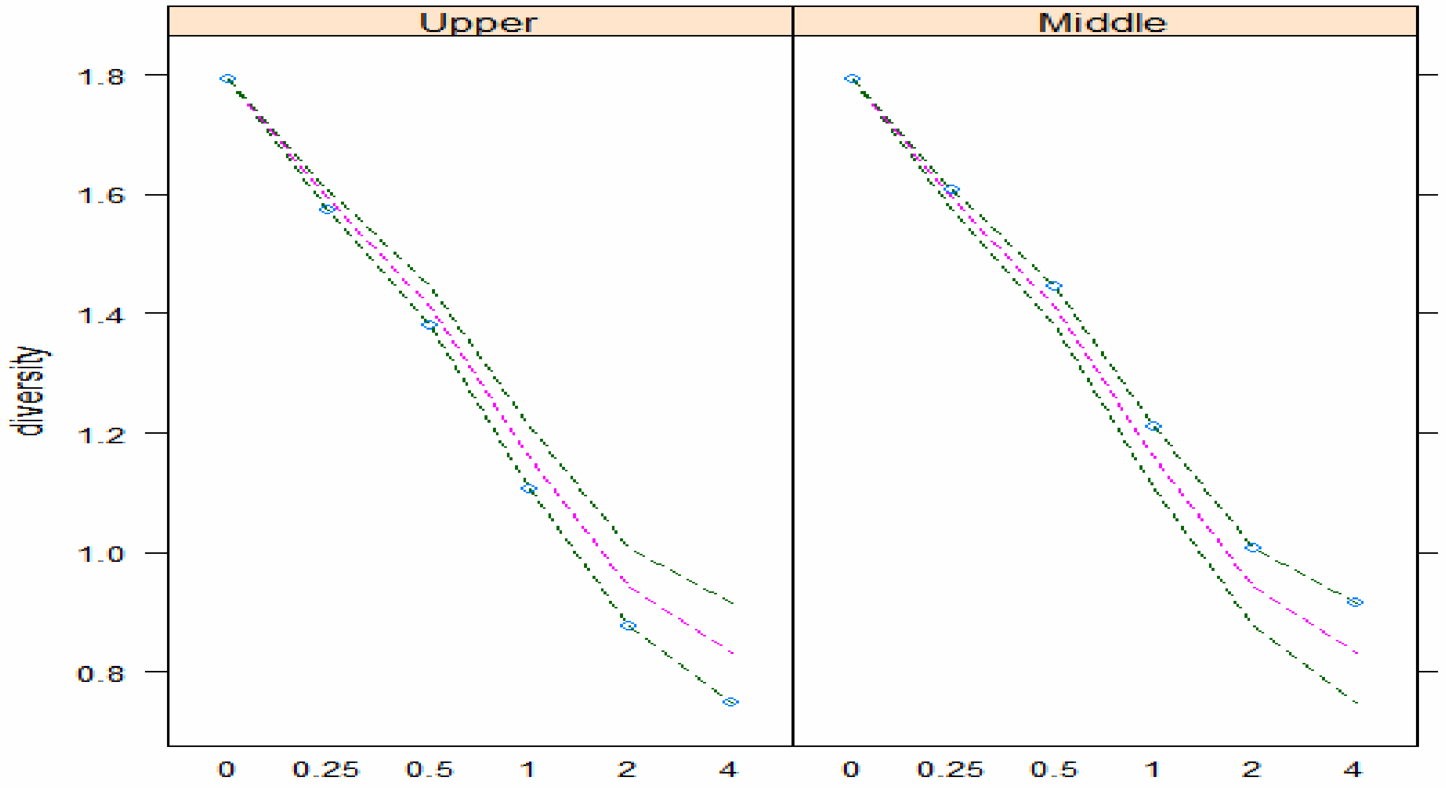
- Number of individuals per family of dinosaur observed

Family	Upper	Middle
Ceratopsidae	50	53
Hadrosauridae	29	51
Hypsilophodontidae	3	2
Tyrannosauridae	3	3
Ornithomimidae	4	8
Saurornithoididae	1	6

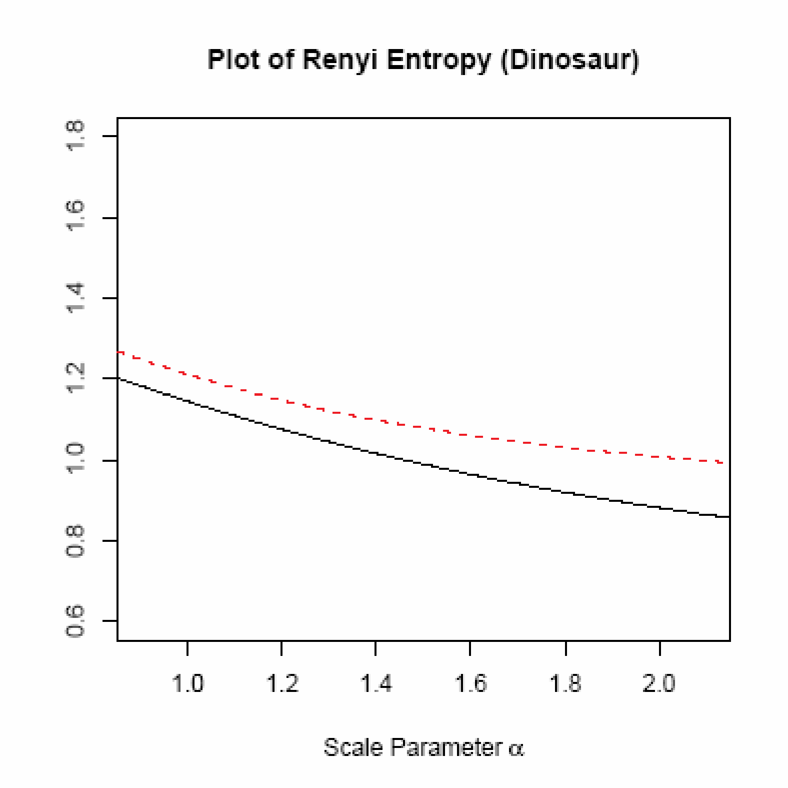
vegan package for Biodiversity

- Here, I will use the **vegan** package to compute Shannon, Simpson, and Renyi diversity and to construct diversity profiles.
- > diversity(Dino) #Shannon index
-    Upper   Middle
- 1.106592 1.210422
- > diversity(Dino,"simpson") #Simpson index
-     Upper   Middle
- 0.5832099 0.6349395

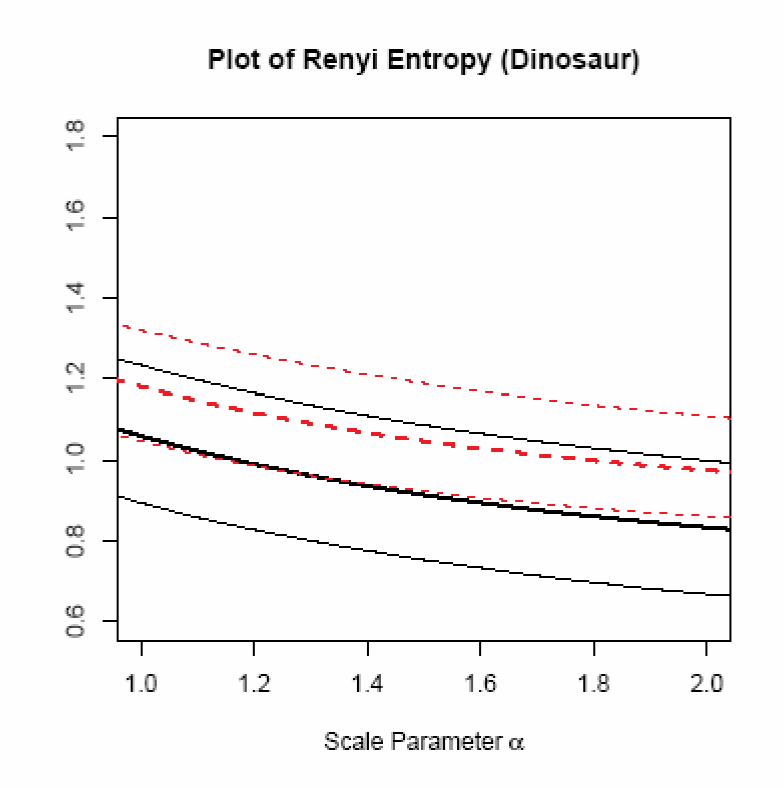
Renyi Profile Plot (vegan)



Renyi profile plot (Mecklin)



Renyi Profile Plots (Mecklin-Bayesian intervals)



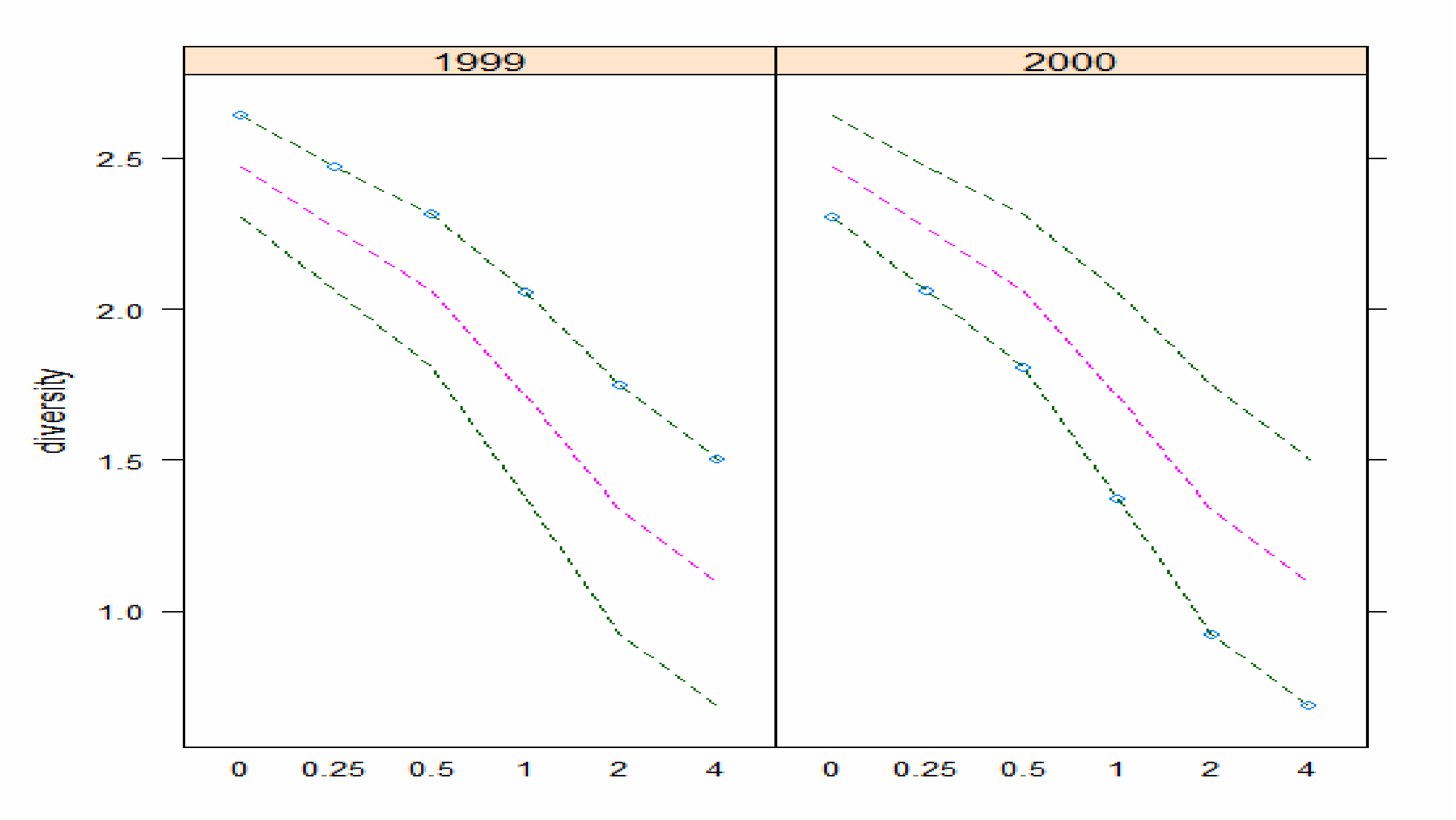
Kentucky Flavor (Mussels data)

- Dr. Jim Sickel and various graduate students assessed the diversity of freshwater mussels at the same location in the Ohio River for several years from 1990-2000.
- Sickel noticed an apparent decrease in mussel diversity between the 1999 and 2000 sampling period.

Mussel Diversity Statistics

```
> diversity(Muss)
     1999      2000
2.055073 1.369266
> diversity(Muss,"simpson")
     1999      2000
0.8254839 0.6011303
```

Renyi Profiles for Mussel Data



BiodiversityR

- Let’s take a quick look at the package I just downloaded.
- I will show you the html help file for this package, and maybe we’ll run through a quick example.
- I’m trying to give you a flavor of the sense of “discovery” I have when I encounter a **R** package for the first time.

Left-Censored Environmental Data

- Dennis Helsel and Lopaka Lee have developed methods for dealing with left-censored environmental data, contained in the **NADA** (Nondetects and Data Analysis) package.
- Helsel had written SAS macros, EXCEL spreadsheets, Minitab programs, and the like in the past, but I guess he got sick of doing all that.

“Nondetects”

- It is very common in studies in environmental chemistry for trace levels of chemical compounds or elements to be reported as “less-than” or “nondetect”.
- The concentration is only known to be between 0 and the lab’s reporting level (RL). A study could have several RLs.
- This is an example of left-censored data.

The problem with “Nondetects”

- Left-censored “less-than” or “nondetect” data complicates the calculation of descriptive statistics and hypothesis testing.
- The traditional “solution” to the nondetect problem has been to either treat the nondetects as missing data or to make a crude imputation (such as ½ the RL) of the missing value(s).
- As Helsel has warned for years, this results in inaccurate statistics and incorrect decisions.

What to do about Nondetects

- Helsel and Lee suggest three methods for estimating the descriptive statistics when there is nondetect data:

1. Kaplan-Meier method (more commonly associated with right-censored data). Basically we estimate the CDF for the data set with a step-function, known as a survival curve. The mean is the area under the survival curve. This method does not assume the distribution of your data.

What to do about Nondetects

1. MLE (maximum likelihood estimation)-typically based on lognormal distribution in this situation. Detects and the proportion falling below RL are used to fit the curve and determine the most "likely" values for the mean & standard deviation. Not suggested for small data sets (an outlier can ruin the results) or when the specified distribution is a poor fit.
2. ROS (Regression on Order Statistics)-an imputation method that instead of assigning all nondetects the same value (which will underestimate the variance), fills in the nondetects based on a probability plot of detects. Multiple RLs can be incorporated

Leah Blackketter’s Study

- Leah was a master’s student in chemistry who recently defended her thesis.
- The premise of her research was determining whether there is a problem with arsenic in groundwater in western Kentucky.

Leah Blackketter’s Study

- A total of n=109 wells in Ballard, Carlisle, and Graves Counties were used. The water samples were tested for arsenic and about 20 other metals and metalloids.
- The reason behind the worry over arsenic is **roxarsone,** an additive in feed used in commercial chicken farms.
- This compound contains arsenic, which is toxic if freed from the molecule.

NADA with an example from Helsel & Lee

- Back to Leah’s study later.
- I used one of the sample data sets that came with the **NADA** package.
- This data set measured the concentration of pyrene at 20 different stations on Puget Sound in Washington state. A total of 8 RLs were present.

Pyrene Study

- The mean and standard deviation of pyrene concentration (which had substantial left-censoring) was made with all three methods:

```
> censtats(Pyrene,PyreneCen)
       n     n.cen  pct.cen
56.00000 11.00000 19.64286
       median      mean       sd
K-M 98.00000 164.2036 393.9509
ROS 90.50000 163.1531 393.1309
MLE 91.64813 133.9142 142.6698
```

Inference with Nondetects

- Helsel and Lee have also discussed a number of parametric, semi-parametric, and non-parametric methods for comparing concentrations between 2 or more populations.
- Much of the commercially available software for “survival analysis”, which uses methods such as Kaplan-Meier, worked with only right-censored or “greater-than” data, which needs to be flipped with “nondetects”.

Good news/Bad news with Leah's data

- First, the bad news. After Leah went through the learning curve of learning **R** and **NADA** and I went through the learning curve of learning the **NADA** methods, we were unable to analyze her arsenic data because she had 100% nondetects.

Good news/Bad news with Leah’s data

- Of course, this is actually good news because it meant that Leah did not find any detectable amounts of arsenic in the groundwater at the sampled wells.
- We were able to use Helsel’s methods with some of the other elements, such as sodium and calcium. The significant difference between aquifers that were found were expected due to the geology of the area.

Ordination-What’s that?

- According to Gauch (1982): "Ordination primarily endeavors to represent sample and species relationships as faithfully as possible in a low-dimensional space".
- Ordination is the arrangement or ‘ordering’ of species and/or sample units along gradients (Palmer).
- If we were better at visualizing many dimensional space, ordination techniques would not be needed. But we aren’t and they are.
- There are dozens of multivariate methods available for ordination. Many of them are based on eigenanalysis.

Duality Diagram Theory

- The French school of ecology is largely based on “duality diagram theory”, implemented in **ade4**.
- **X** is a data table (matrix) with n rows (individuals) and p columns (variables).
- The columns could represent species abundances and/or the values of quantitative/qualitative environmental variables.

Duality Diagram Theory

- We will “handwave” a ton of linear algebra away.
- Two other matrices, **Q** and **D**, are needed. (Warning: Everyone uses different notation in the ordination world!)
- **Q** is a p by p positive symmetric matrix (distance between n individuals)
- **D** is a n by n positive symmetric matrix (relationship between p variables)
- Basically, the various ordination methods seek a low dimensional hyperspace that represents individuals as closely as possible to the original space.

Principal Components Analysis

- Geometrically, PCA is a rigid rotation of the original data matrix, and can be defined as a projection of samples onto a new set of axes, such that the maximum variance is projected or "extracted" along the first axis, the maximum variation uncorrelated with axis 1 is projected on the second axis, the maximum variation uncorrelated with the first and second axis is projected on the third axis, etc. (Palmer)

Eigenvalue based ordination methods

- An eigenanalysis is performed on a square, symmetric matrix derived from the data matrix
- Each ordination axis is an eigenvector, and is associated with an eigenvalue. The coordinates for the ith sample along a given axis is the ith element of the axis' eigenvector.
- Axes are ranked by their eigenvalues. Thus, the first axis has the highest eigenvalue, the second axis has the second highest eigenvalue, etc.
- Eigenvalues have mathematical meaning that can aid in interpretation. In principal components analysis, eigenvalues are 'variance extracted'
- Similar to "factor analysis" which is popular with social scientists.

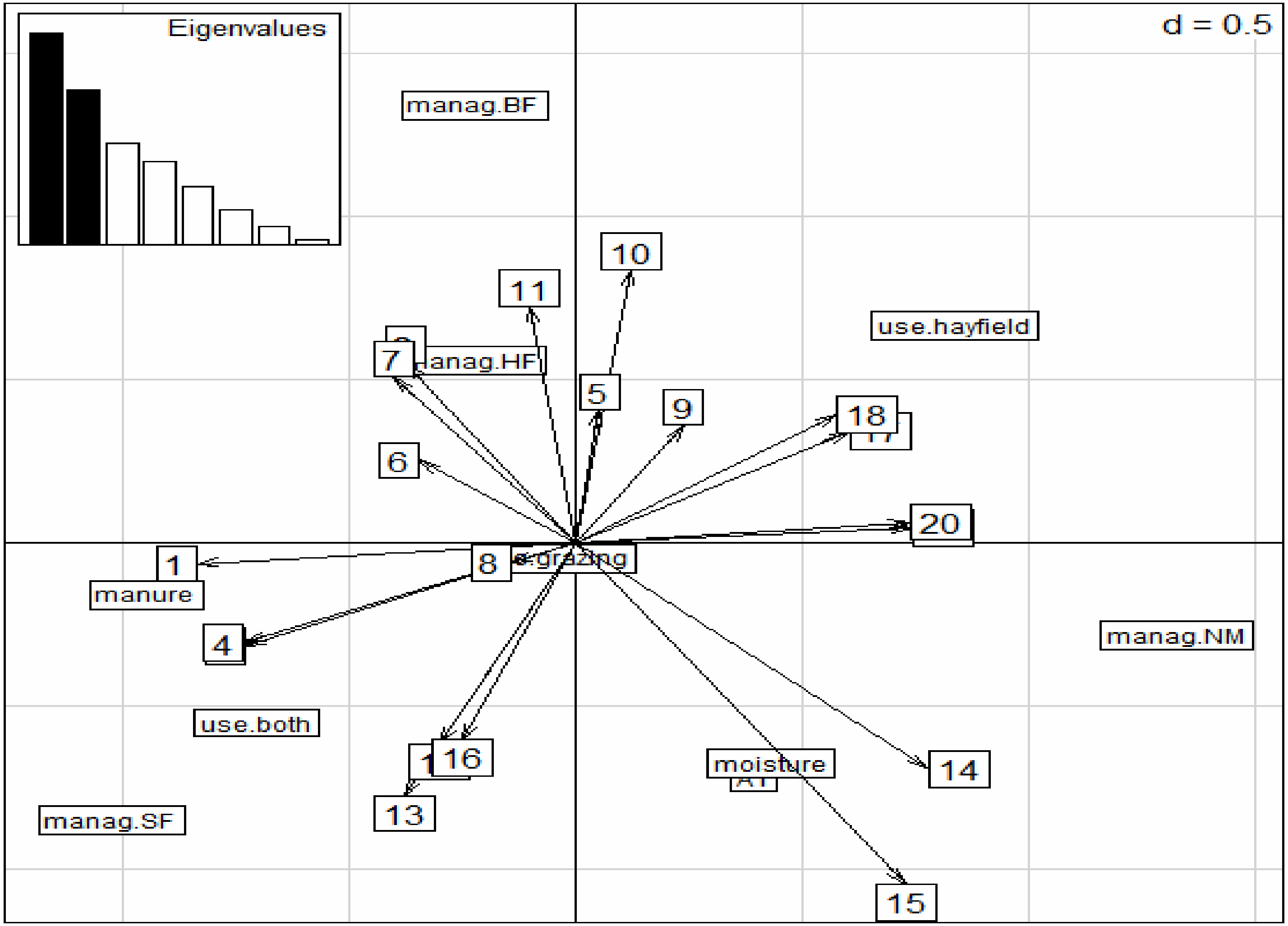
Hill-Smith Variation on PCA

- The method illustrated in the Bray/Duffour article is a variation on principal components analysis that allows for both quantitative and categorical variables.
- This will be illustrated with the “Dune Meadows” data set, with 20 sites and 5 variables.

Dune Meadows

- A1-thickness of the A1 horizon
- Moisture-moisture content of soil
- Manure-quantity of manure applied
- Use-categorical factor with 3 levels (hayfield, grazing, both)
- Management-categorical factor with 4 levels (standard farming, biological farming, hobby farming, nature conservation management)

Biplot of Dune Meadow Data



Interpretation of Biplot

- If you are lucky, the first couple of principal components will make biological sense.
- Axis 1 discriminates between sites with high manure use (standard farming) versus the conserved sites.
- Axis 2 seperates sites with high moisture & A1 horizon from the drier sites.

Eigenanalysis results

- eigen values: 2.542 1.858 1.231 0.9899 0.6927 ... (sum=8)

```
 Axis1      Axis2
1  -2.5387568 -0.1678061
2  -1.1346547  1.4264458
3  -2.2103007 -0.8209185
Etc.
```

```
               Comp1       Comp2
A1             0.39417183 -0.71676249
moisture       0.43123610 -0.67503957
manure        -0.94661847 -0.16039063
use.hayfield   0.83813897  0.66356117
use.both      -0.70491509 -0.55114781
use.grazing   -0.04553042 -0.04714914
manag.BF      -0.22373287  1.33796380
manag.HF      -0.22756390  0.55664577
manag.NM       1.32365715 -0.28364213
manag.SF      -1.02215413 -0.84921124
```

Ordination resources

- The Bray-Duffour article from *Journal of Statistical Software* (most articles by these researchers & colleagues are in French)
- *Numerical Ecology* by Legendre & Legendre
- Michael Palmer’s webpage http://ordination.okstate.edu

Other uses of **R**

- Bayesian analysis (Gibbs sampler, WinBUGS) via the packages **BRugs** and **R2WinBUGS** (I’ve used BRugs as a substitute to WinBUGS and it seems to work fine)
- Bootstrapping (**boot**, **bootstrap**, etc.)
- Lots more I don’t have time to mention or even know about

Inventing and **R**e-inventing the Wheel

- If what you need isn't in one of the canned packages, you can always write your own programs (i.e. no one has invented your wheel yet).
- Some of these home-brewed programs end up being cleaned up and turned into an **R** package (so no one else has to re-invent your wheel).

Conclusion

- I failed to teach you how to program in **R** today. Luckily, that wasn’t my goal.
- Something to keep in mind-you may very well have a data analysis need someday that will be best met by using **R**
- Maybe someone should spend his summer writing an **R** package for “Fluctuating Asymmetry”

Are you ready to convert to **R**-ism?

- The **R** project homepage is http://cran.r-project.org
- A lot of good tips on how to do your bread-and-butter statistical analyses with **R** is at http://www.statmethods.net (this website was created by a long-time user of SAS who had some of the same difficulties in getting used to **R** that I did!)
- Did I remember to mention that it is **free**?
- My email is christopher.mecklin@murraystate.edu