

Week 1 Exercise: Basic R

Z620: Quantitative Biodiversity, Indiana University

January 16, 2015

OVERVIEW

This exercise will introduce some of the basic features of the R computing environment. We will briefly cover operators, data types, and simple commands that will be useful for you during the course and beyond. In addition to using R's base package, we will also use contributed packages, which together will allow us to visualize data and perform relatively simple statistics (e.g., linear regression and ANOVA).

1) HOW WE WILL BE USING R AND OTHER TOOLS

During the course, we will use [RStudio](#), which is a relatively user-friendly integrated development environment (IDE) that allows R to interface with other tools. For example, you are currently working in an [R Markdown](#) document. Markdown is a simple formatting syntax for authoring HTML, PDF, and other documents.

We will also use a tool called [knitr](#), which is a package that generates reports from R script and Markdown text. For example, when you click the **Knit PDF** button in the scripting window of Rstudio, a document will be generated that includes [LaTeX](#) typesetting as well as the output of any embedded R code.

If there are errors in your markdown document, however, you will not be able to knit a PDF file. Assignments in this class will require that you successfully create a Markdown-generated PDF using knitr; you will then need to **push** this document to the course **repository** hosted on [IU's GitHub](#) and generate a **pull request**.

2) SETTING YOUR WORKING DIRECTORY

A good first step when you sit down to work in R is to clear your working directory of any variables from your workspace:

```
rm(list=ls()) # removes all variables from your workspace
```

Now we want to set the working directory. This is where your R script and output will be saved. It's also a logical place to put data files that you plan to import into R. The following command will return your current working directory:

```
getwd()
```

```
## [1] "C:/Users/Mario Muscarella/GitHub/QuantitativeBiodiversity/Assignments/Week1"
```

Use the following command to change your directory (but note that you will need to modify to reflect *your* actual directory):

```
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/Week1")
```

3) USING R AS A CALCULATOR

R is capable of performing various calculations using simple operators and built-in **functions**

- *addition*:

```
1 + 3
```

```
## [1] 4
```

- *subtraction*:

```
3 - 1
```

```
## [1] 2
```

- *multiplication* (with an exponent):

```
3 * 10^2
```

```
## [1] 300
```

- *division* (using a built-in constant; pi):

```
10 / pi
```

```
## [1] 3.183099
```

- *trigonometry* with a simple built-in **function** (i.e., *sin*) that takes an **argument** (i.e., '4'):

```
sin(4)
```

```
## [1] -0.7568025
```

- *logarithms* (another example of functions and arguments)

```
log10(100)
```

```
## [1] 2
```

```
log(100)
```

```
## [1] 4.60517
```

4) ASSIGNING VARIABLES

In R, you will often find it useful and necessary to assign values to a variable or **object**. Generally speaking, in R, it's best to use <- rather than = as an assignment operator.

```
a <- 10
b <- a + 20
```

What is the value of b?

Now let's reassign a new value to a:

```
a <- 200
```

Now, what is the value of 'b'? What's going on?

R held onto the original value of 'a' that was used when assigning values to 'b'. You can correct this using the `rm` function, which removes objects from your R **environment**.

```
rm("b")
```

What happens if we reassign b now?

```
b <- a + 20
```

Sometimes it's good practice to clear all variables from your R environment (e.g., you've been working on multiple projects). This can be done in a couple of ways. For example, you can just click **clear** in the Environment window of R Studio. The same procedure can be performed at the command line. To do this, you can use the `ls` function to view a list of all the objects in the R environment:

```
ls()
```

```
## [1] "a" "b"
```

You can now clear all of the stored variables from R's memory (using two functions: `rm` and `ls`) as we did above prior to setting our working directory:

```
rm(list=ls())
```

5) WORKING WITH VECTORS

Basic Features Of Vectors

Vectors are the fundamental data type in R. Often, vectors are just a collection of data of a similar type, either numeric (e.g., 17.5), integer (e.g., 2), or character (e.g., "low"). The simplest type of vector is a single value, sometimes referred to as a **scalar** in other programming languages:

```
w <- 5
```

We can create longer one-dimensional vectors in R like this:

```
x <- c(2, 3, 6, w, w + 7, 12, 14)
```

What is the function `c()` that we just used to create a vector? To answer this question, try typing `help()` function at the command line. Let's try it out:

```
help(c)
```

```
## starting httpd help server ... done
```

What happens when you multiply a vector by a “scalar”?

```
y <- w * x
```

What happens when you multiply two vectors of the same length?

```
z <- x * y
```

You may need to reference a specific **element** in a vector. We will do this using the square brackets. In this case, the number inside of the square brackets tells R that we want call the second element of vector **z**:

```
z[2]
```

```
## [1] 45
```

You can also reference multiple elements in a vector:

```
z[2:5]
```

```
## [1] 45 180 125 720
```

In some instances, you may want to change the value of an element in a vector. Here’s how you can substitute a new value for the second element of **z**:

```
z[2] <- 583
```

Summary Statistics Of Vectors

It’s pretty easy to perform summary statistics on a vector using the built-in fuctions of R:

```
max(z)      # maximum
```

```
## [1] 980
```

```
min(z)      # minimum
```

```
## [1] 20
```

```
sum(z)      # sum
```

```
## [1] 3328
```

```
mean(z)    # mean
```

```
## [1] 475.4286
```

```
median(z)  # median
```

```
## [1] 583
```

```
var(z)     # variance
```

```
## [1] 133881.3
```

```
sd(z)      # standard deviation
```

```
## [1] 365.8979
```

What happens when you take the standard error of the mean (**sem**) of *z*?

The standard error of the mean is defined as $sem = \frac{sd(x)}{\sqrt{n}}$. This function does not exist in the base package of R. Therefore, you need to write your own function. Let's give it a try:

```
sem <- function(x, ...){  
  sd(x)/sqrt(length(x))  
}
```

There are number of functions inside of **sem**. Take a moment to think about and describe what is going on here. Use the **sem** function you just created on the vector *y* from above

Often, datasets have missing values (designated as 'NA' in R):

```
i <- c(2, 3, 9, NA, 120, 33, 7, 44.5)
```

What happens when you apply your **sem** function to vector *i*? This is a problem!

We should modify the original **sem** function so it removes NAs right from the start. We can do this by using the **na.rm** function within the **sd** function and the **na.omit** function within **length** function.

```
sem <- function(x){  
  sd(x, na.rm = TRUE)/sqrt(length(na.omit(x)))  
}
```

Now run **sem** on *i*.

5) WORKING WITH MATRICES

Matrices are another data type in R. They are just two-dimensional vectors containing data of the same type (e.g., numeric, integer, character). Therefore, much of what we just discussed about vectors translates directly into dealing with matrices.

Making A Matrix

There are three common ways to create a matrix in R.

Approach 1 is to combine (or concatenate) two or more vectors. Let's start by creating a one-dimensional vector using a new function `rnorm` based on information contained in vector `z` above.

```
j <- c(rnorm(length(z), mean = z))
```

What does the `rnorm` function do? What are arguments specifying? Remember your friend `help` or type `?rnorm`.

Now we will use the function `cbind` to create a matrix by combining the two one-dimensional vectors:

```
k <- cbind(z, j)
```

Use the `help` function to learn about `cbind`. Use the `dim` function to describe the matrix you just created. What did you learn from this?

Approach 2 to making a matrix is to use the `matrix` function along with arguments that specify the number of rows (`nrow`) and columns (`ncol`):

```
l <- matrix(c(2, 4, 3, 1, 5, 7), nrow = 3, ncol = 2)
```

Approach 3 to making a matrix is to import or *load a dataset* from your working directory:

```
m <- as.matrix(read.table("matrix.txt", sep = "\t", header = FALSE))
```

In this case, we're reading in a tab-delimited file. The name of your file must be in quotes, and you need to specify tab-delimited file type using the `sep` argument. The `header` argument tells R whether or not the names of the variables are contained in the first line; in the current example, they are not.

Often, when handling datasets, we want to be able to **transpose** a matrix. This is an easy operation in R that uses the `t` function:

```
n <- t(m)
```

Confirm the transposition using the `dim` function.

Indexing A Matrix

Frequently, you will need to **index** or retrieve a certain portion of a matrix. As with the vector example above, we will use the square brackets to retrieve data from a matrix. Inside the square brackets, there are now two subscripts corresponding to the rows and columns, respectively, of the matrix.

The following code will create a new matrix (`n'`) based on the first three rows of matrix (`m'`):

```
n <- m[1:3, ]
```

Or maybe you want the first two columns of a matrix instead:

```
n <- m[, 1:2]
```

Or perhaps you want non-sequential columns of a matrix. How do we do that? It's easy when you understand how to reference data within a matrix:

```
n <- m[, c(1:2, 5)]
```

Describe what we just did in the last indexing operation.

6) BASIC DATA VISUALIZATION AND STATISTICAL ANALYSIS

Load Zooplankton Dataset

In the following exercise, we will use a dataset from [Lennon et al. \(2003\)](#), which looked at zooplankton community assembly along an experimental nutrient gradient. Inorganic nitrogen and phosphorus were added to mesocosms for six weeks at three different levels (low, medium, and high), but we also directly measured nutrient concentrations in the mesocosms. So we have categorical and continuous predictors that we're going to use to help explain variation in zooplankton biomass.

The first thing we're going to do is load the data:

```
meso <- read.table("zoop_nuts.txt", sep = "\t", header = TRUE)
```

Let's use the `str` function to look at the structure of the data.

```
str(meso)
```

```
## 'data.frame': 24 obs. of 8 variables:
## $ TANK: int 34 14 23 16 21 5 25 27 30 28 ...
## $ NUTS: Factor w/ 3 levels "H","L","M": 2 2 2 2 2 2 2 2 3 3 ...
## $ TP : num 20.3 25.6 14.2 39.1 20.1 ...
## $ TN : num 720 750 610 761 570 ...
## $ SRP : num 4.02 1.56 4.97 2.89 5.11 4.68 5 0.1 7.9 3.92 ...
## $ TIN : num 131.6 141.1 107.7 71.3 80.4 ...
## $ CHLA: num 1.52 4 0.61 0.53 1.44 1.19 0.37 0.72 6.93 0.94 ...
## $ ZP : num 1.781 0.409 1.201 3.36 0.733 ...
```

How does this dataset differ from the `m` dataset above? We're now dealing with a new type of **data structure**. Specifically, the `meso` dataset is a **data frame** since it has a combination of numeric and character data. (Remember, **matrices** and **vectors** are only comprised of a *single* data type.)

Here is a description of the column headers:

- TANK = unique mesocosm identifier
- NUTS = categorical nutrient treatment: "L" = low, "M" = medium, "H" = high
- TP = total phosphorus concentration ($\mu\text{g/L}$)
- TN = total nitrogen concentration ($\mu\text{g/L}$)
- SRP = soluble reactive phosphorus concentration ($\mu\text{g/L}$)
- TIN = total inorganic nutrient concentration ($\mu\text{g/L}$)
- CHLA = chlorophyll *a* concentration (proxy for algal biomass; $\mu\text{g/L}$)
- ZP = zooplankton biomass (mg/L)

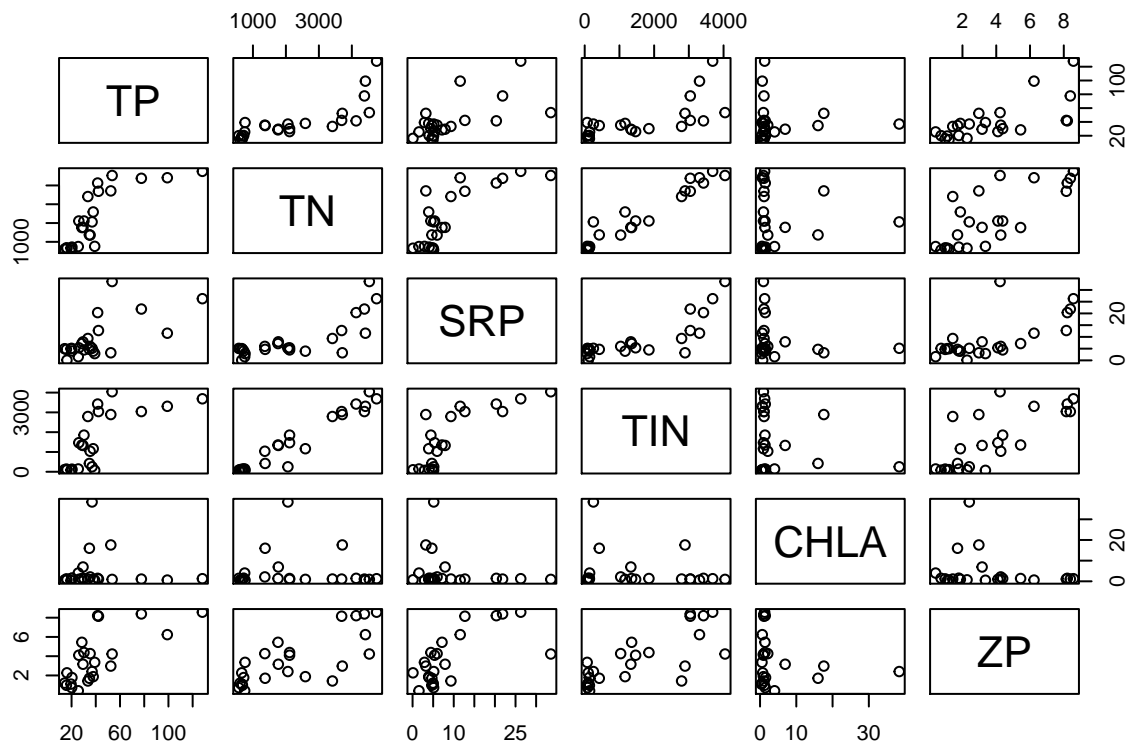
Correlation

A common step in data exploration is to look at correlations among variables. Before we do this, let's **index** our numerical (continuous) data in the 'meso' dataframe. (Correlations typically don't work well on categorical data.)

```
meso.num <- meso[,3:8]
```

We can conveniently visualize pairwise **bi-plots** of the data using the following command:

```
pairs(meso.num)
```



Now let's conduct a simple **Pearson's correlation** analysis with the `thecor()` function.

```
cor1 <- cor(meso.num)
```

Describe what you found from the visualization and correlation analysis above?

want to have some way to "grey out" an area where students will be expected to fill in answers {JTL}

Loading Contributed Packages

The base package in R won't always meet all of our needs. This is why there are > 6,000 **contributed packages** that have been developed for R. This may seem overwhelming, but it also means that there are tools (and web support) for just about any problem you can think of.

When using one of the contributed packages, the first thing we need to do is **install** them along with their dependencies (other required packages). We're going to start out by using the *psych* package. The *psych* package has many features, but we're going to use it specifically for the `corr.test` function. This function generates p-values for each pairwise correlation. (For whatever reason, the `cor` function in the **base** package of R does not generate p-values.)

You can load an R package and its dependencies using the `require()` function. With the following string of commands, if the package is not found using `require()`, R will use the `install.packages()` function followed by `require()`:

```
require("psych") || install.packages("psych");require("psych")
```

```
## Loading required package: psych
```

```
## [1] TRUE
```

Now, let's look at the correlations among variables and assess whether they are significant:

```
cor2 <- corr.test(meso.num, method = "pearson", adjust = "BH")
print(cor2, digits = 3)
```

```
## Call:corr.test(x = meso.num, method = "pearson", adjust = "BH")
## Correlation matrix
##      TP      TN      SRP      TIN      CHLA      ZP
## TP   1.000  0.787  0.654  0.717 -0.017  0.697
## TN   0.787  1.000  0.784  0.969 -0.004  0.756
## SRP  0.654  0.784  1.000  0.801 -0.189  0.676
## TIN  0.717  0.969  0.801  1.000 -0.157  0.761
## CHLA -0.017 -0.004 -0.189 -0.157  1.000 -0.183
## ZP   0.697  0.756  0.676  0.761 -0.183  1.000
## Sample Size
## [1] 24
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##      TP      TN      SRP      TIN      CHLA      ZP
## TP   0.000  0.000  0.001  0.000  0.983  0.000
## TN   0.000  0.000  0.000  0.000  0.983  0.000
## SRP  0.001  0.000  0.000  0.000  0.491  0.000
## TIN  0.000  0.000  0.000  0.000  0.536  0.000
## CHLA 0.938  0.983  0.376  0.464  0.000  0.491
## ZP   0.000  0.000  0.000  0.000  0.393  0.000
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

Notes on corr.test:

- for rank-based correlations (i.e., non-parametric), use `method = "kendall"` or `"spearman"`. Give it a try!
- the `adjust = "BH"` statement supplies the Benjamini & Hochberg-corrected **p-values** in the upper right diagonal of the square matrix; the uncorrected p-values are below the diagonal. This process corrects for **false discovery rate**, which arises when making multiple comparisons.

Describe what you learned from `corr.test` and the notes:

```
###fill in blank###
```

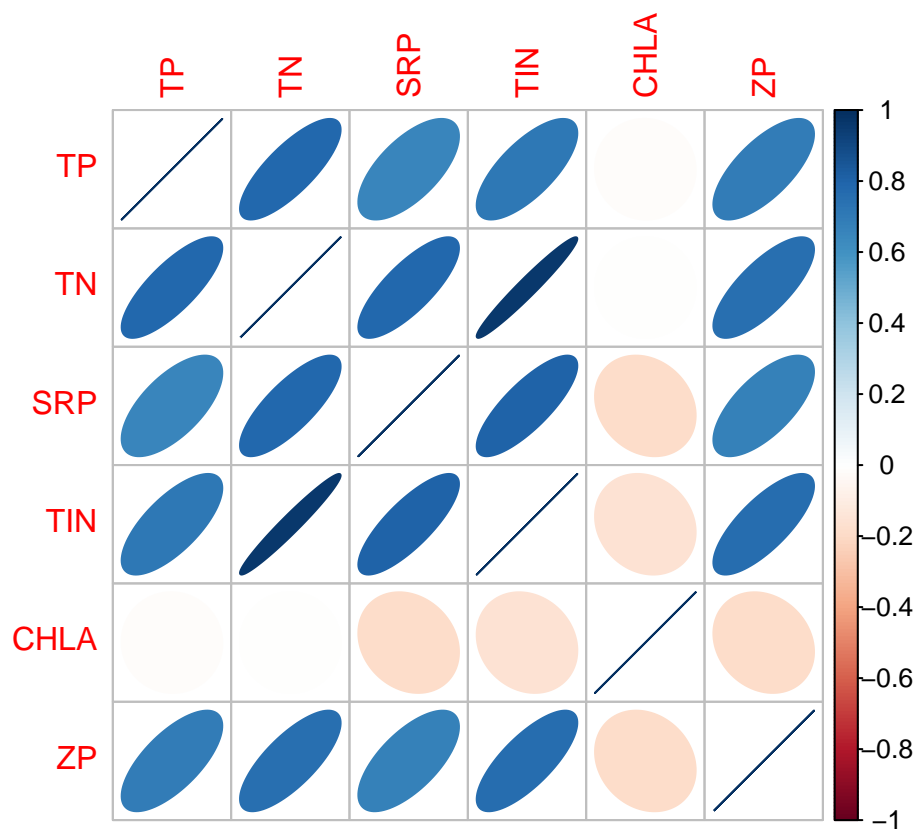
Now, Let's load another package that will let us visualize the sign and strength of the correlations:

```
require("corrplot")||install.packages("corrplot");require("corrplot")
```

```
## Loading required package: corrplot
```

```
## [1] TRUE
```

```
corrplot(cor1, method = "ellipse")
```



Linear Regression

It seems that total nitrogen (TN) is a fairly good predictor of zooplankton biomass (ZP) and this is something that we directly manipulated. This gives us license to conduct a linear regression analysis. We can do this in R using the `lm` function:

```
fitreg <- lm(ZP ~ TN, data = meso)
```

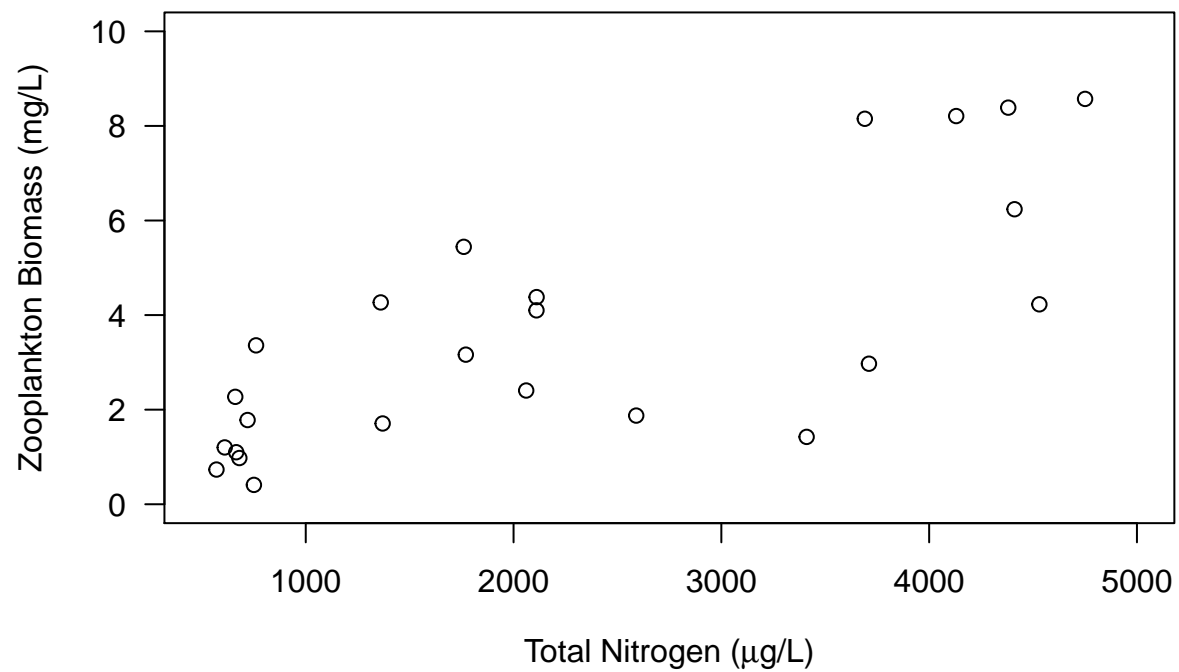
Let's examine the output of the regression model:

```
summary(fitreg)
```

```
##
## Call:
## lm(formula = ZP ~ TN, data = meso)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.7690 -0.8491 -0.0709  1.6238  2.5888
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6977712  0.6496312   1.074    0.294
## TN          0.0013181  0.0002431   5.421 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.75 on 22 degrees of freedom
## Multiple R-squared:  0.5719, Adjusted R-squared:  0.5525
## F-statistic: 29.39 on 1 and 22 DF,  p-value: 1.911e-05
```

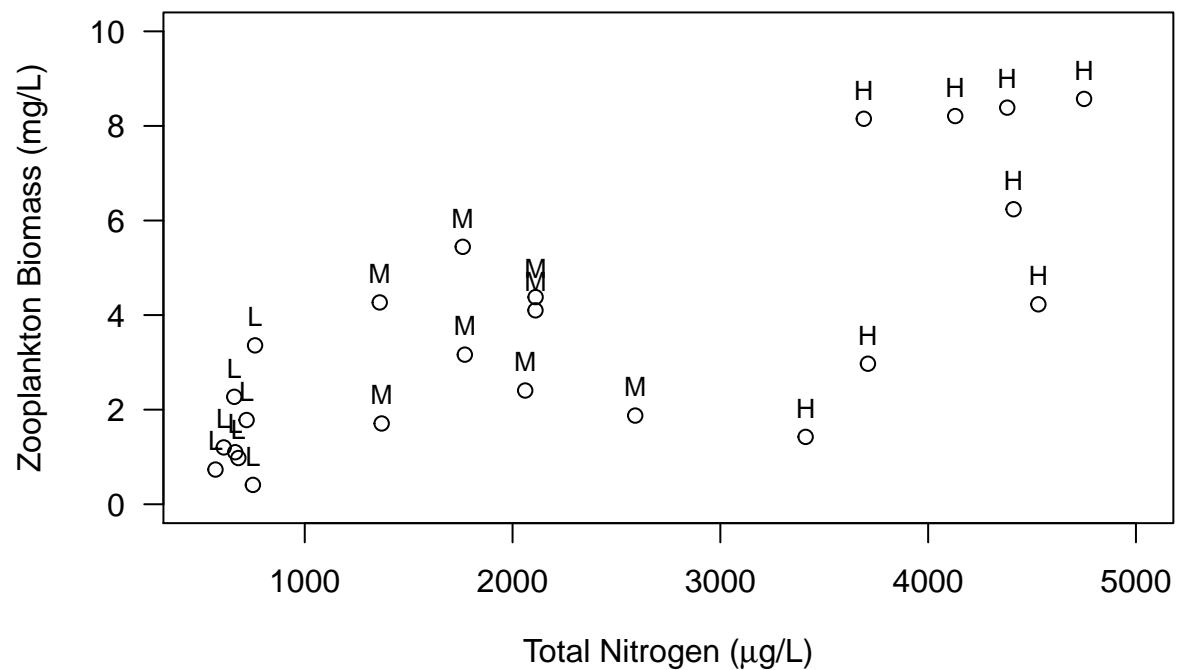
Now, let's look at a plot of the data used in the regression analysis:

```
plot(meso$TN, meso$ZP, ylim = c(0, 10), xlim = c(500, 5000),
     xlab = expression(paste("Total Nitrogen (", mu, "g/L)")),
     ylab = "Zooplankton Biomass (mg/L)", las = 1)
```



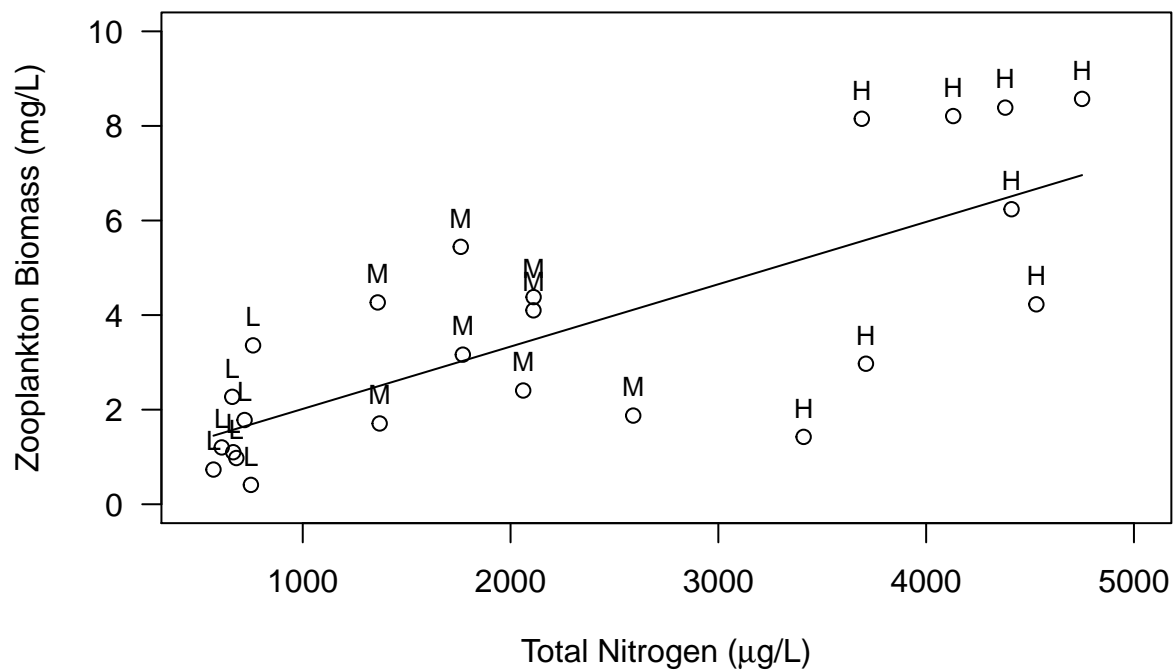
We can add some text to the plot to visualize the categorical nutrient treatments:

```
text(meso$TN, meso$ZP,meso$NUTS,pos=3,cex=0.8)
```



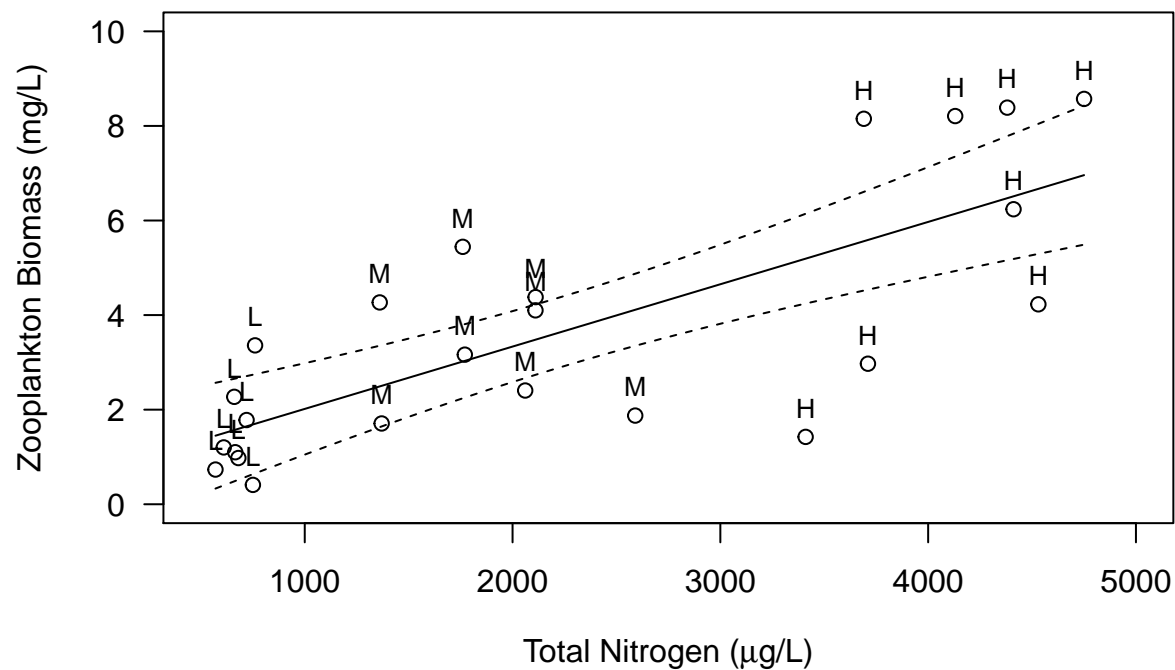
To add the regression line, the first thing we need to do is identify a range of x-values and then generate the corresponding **predicted values** from our regression model:

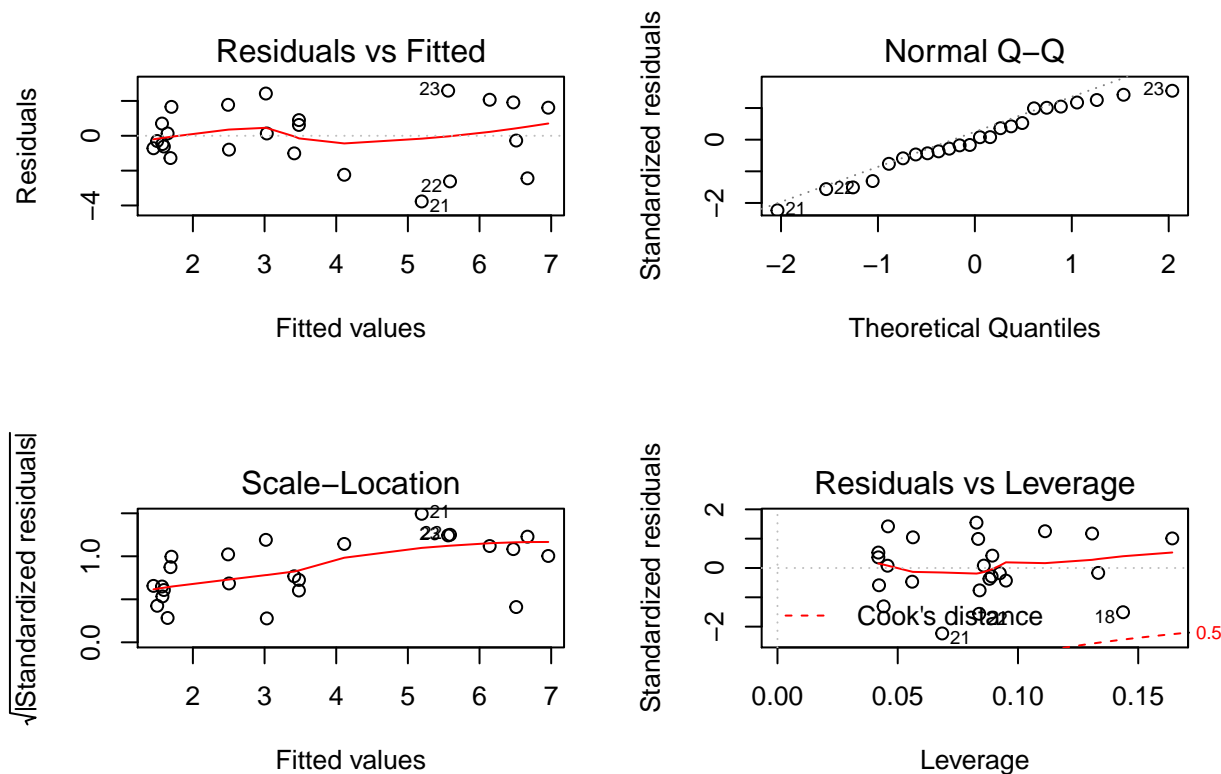
```
text(meso$TN, meso$ZP, meso$NUTS, pos=3, cex=0.8)
newTN <- seq(min(meso$TN), max(meso$TN), 10)
regline <- predict(fitreg, newdata = data.frame(TN = newTN))
lines(newTN, regline)
```



Now let's create and plot the **95% confidence intervals** using the same procedure as above, that is, use 'newTN' to generate corresponding confidence intervals from our regression model:

```
text(meso$TN, meso$ZP, meso$NUTS, pos=3, cex=0.8)
newTN <- seq(min(meso$TN), max(meso$TN), 10)
regline <- predict(fitreg, newdata = data.frame(TN = newTN))
lines(newTN, regline)
conf95 <- predict(fitreg, newdata = data.frame(TN = newTN),
                  interval = c("confidence"), level = 0.95, type = "response")
matlines(newTN, conf95[, c("lwr", "upr")], type="l", lty = 2, lwd = 1, col = "black")
```





+ Upper left: is there a random distribution of the residuals around zero (horizontal line)? + Upper right: is there a reasonably linear relationship between standardized residuals and theoretical quantiles? Try `help(qqplot)` + Bottom left: again, looking for a random distribution of $\sqrt{\text{standardized residuals}}$ + Bottom right: leverage indicates the influence of a point; contours correspond with Cook's distance, where values > 1 are "suspicious"

ANALYSIS OF VARIANCE (ANOVA)

We also have the option of looking at the relationship between zooplankton and nutrients where the manipulation is treated categorically (low, medium, high). First, let's order the categorical nutrient treatments from low to high (R's default is to order alphabetically):

```
NUTS <- factor(meso$NUTS, levels = c('L', 'M', 'H'))
```

Before plotting, we need to calculate the means and standard errors for zooplankton biomass in our nutrient treatments. We are going to use an important function called `tapply`. This allows us to apply a function (e.g., `mean`) to a vector (e.g., `ZP`) based on information in another column (e.g., nutrient treatment).

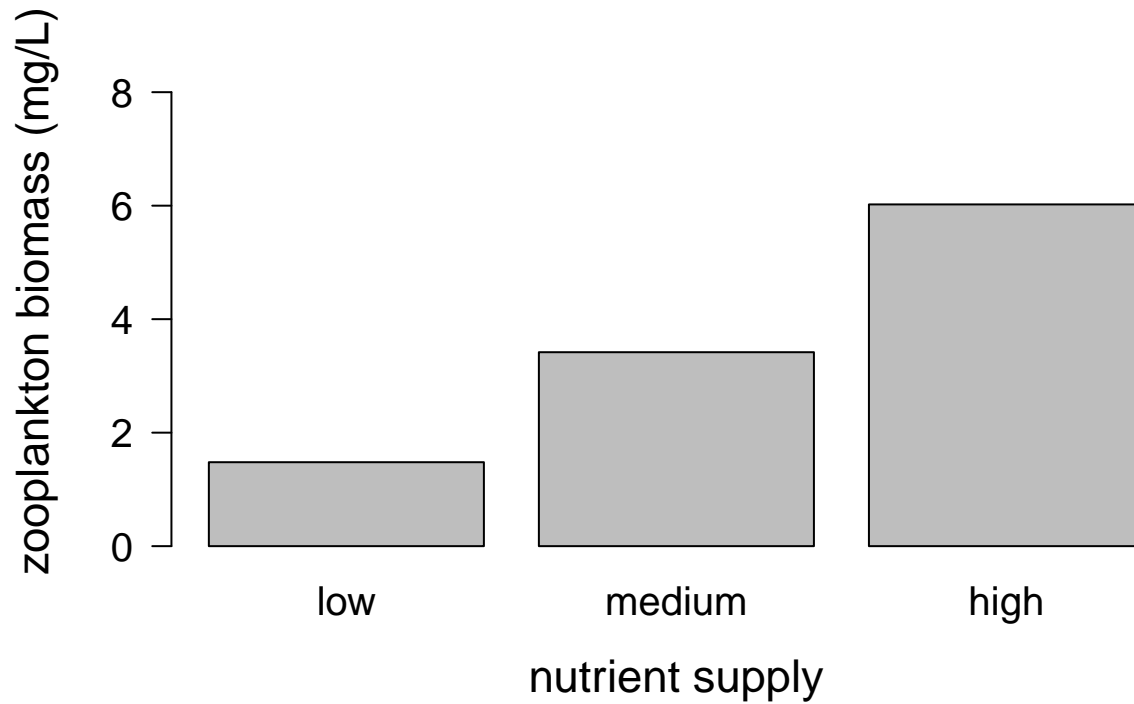
```
zp.means <- tapply(meso$ZP, NUTS, mean)
zp.sem <- tapply(meso$ZP, NUTS, sem)
```

Now let's make the barplot:

```
bp <- barplot(zp.means, ylim = c(0, round(max(meso$ZP), digits = 0)),
             pch = 15, cex = 1.25, las = 1, cex.lab = 1.4, cex.axis = 1.25,
```

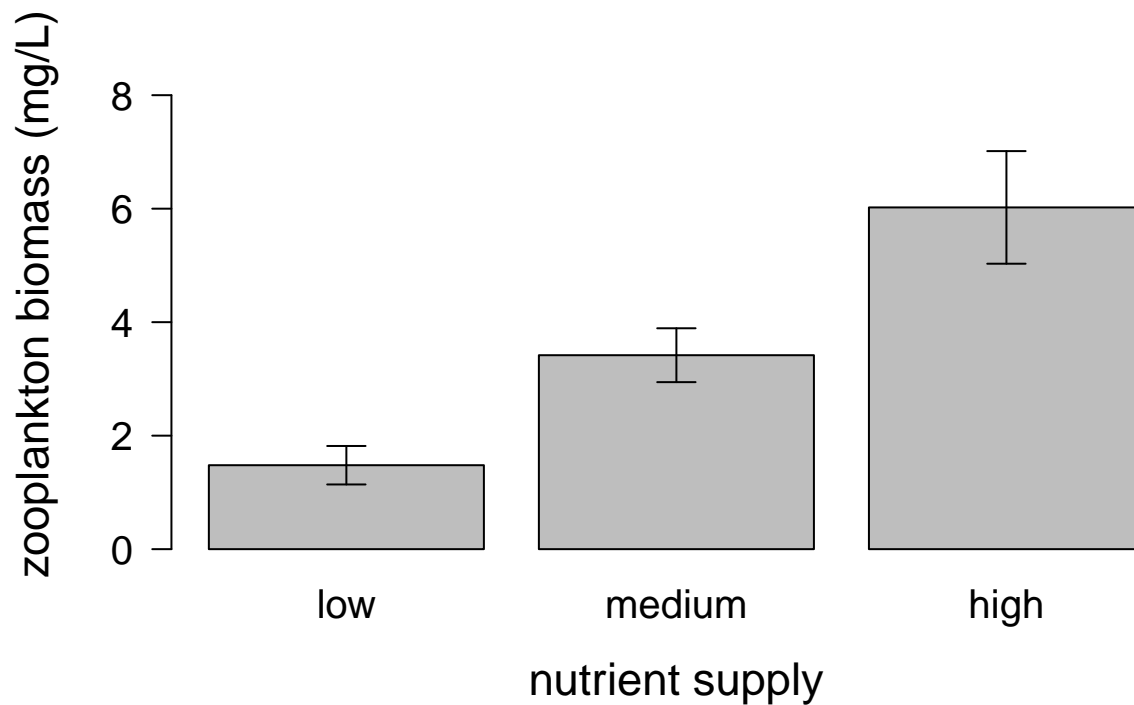


```
xlab = "nutrient supply",  
ylab = "zooplankton biomass (mg/L)",  
names.arg = c("low", "medium", "high"))
```



We need to add the error bars (\pm sem) “manually” as follows:

```
arrows(x0 = bp, y0 = zp.means, y1 = zp.means - zp.sem, angle = 90,  
       length=0.1, lwd = 1)  
arrows(x0 = bp, y0 = zp.means, y1 = zp.means + zp.sem, angle = 90,  
       length=0.1, lwd = 1)
```



We can conduct a one-way analysis of variance (ANOVA) as follows:

```
fitanova <- aov(ZP ~ NUTS, data = meso)
```

Let's look at the output in more detail (just as we did with regression):

```
summary(fitanova)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## NUTS           2  83.15   41.58    11.77 0.000372 ***
## Residuals     21  74.16    3.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we can conduct a post-hoc comparison of treatments using Tukey's HSD (Honest Significant Differences). This will tell us whether or not there are differences ($\alpha = 0.05$) among pairs of the three nutrient treatments

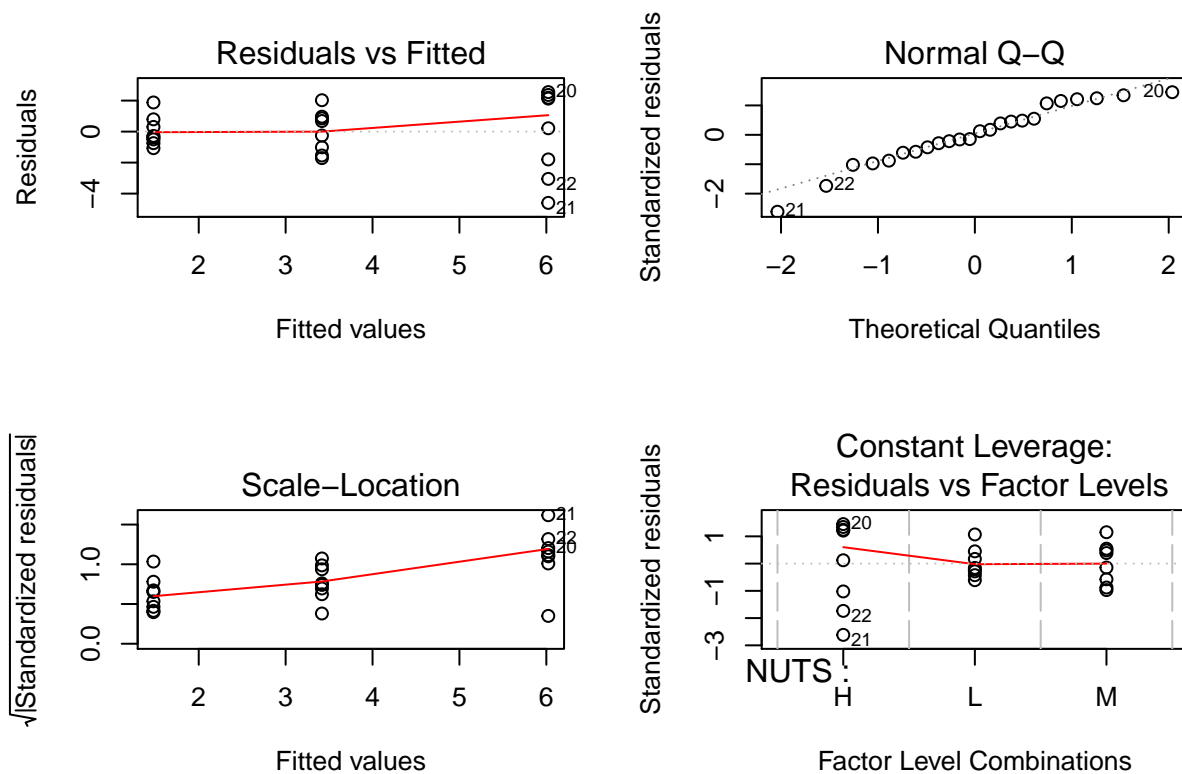
```
TukeyHSD(fitanova)
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = ZP ~ NUTS, data = meso)
```

```
##
## $NUTS
##      diff      lwr      upr      p adj
## L-H -4.543175 -6.9115094 -2.1748406 0.0002512
## M-H -2.604550 -4.9728844 -0.2362156 0.0294932
## M-L  1.938625 -0.4297094  4.3069594 0.1220246
```

Just like the regression analysis above, it's good to look at the residuals:

```
par(mfrow = c(2, 2), mar = c(5.1,4.1,4.1,2.1))
plot(fitanova)
```



HOMEWORK

- 1) Complete this the entire exercise, but
- 2) Redo the linear regression and ANOVA with log10-transformed zooplankton biomass data (often, log-transformations help with meeting assumptions of normality and equal variance). Do you think this is a better analysis?
- 3) Use knitr to create a pdf of your completed exercise, push it to GitHub, and create a pull request.