**EXERCISE 3. LINEAR REGRESSION**

#####################################

# 1. DATA NORMALITY

#####################################

# To decide which statistics to use, we need to know if the richness

# normally distributed?

# Have a look at the data by plotting a histogram

hist(mgr$GenusRich)

# This shows us we have lots of sites with low richness and few with high richness

# So a square root transformation will help....

sqrt.rich <- sqrt(mgr$GenusRich)

hist(sqrt.rich)

# much better!

# a column already exists in the dataframe with square root of genus

#####################################

# 2. CORRELATION BETWEEN 2 VARIABLES

#####################################

# Correlation - two variables

# Is minimum sea surface temperature (SST) correlated with maximum SST?

cor.test(mgr$SST.MIN, mgr$SST.MAX, method="pearson")

# can you see the correlation if you plot these data?

plot(mgr$SST.MIN, mgr$SST.MAX)

# Get a correlation matrix i.e. correlations of multiple variables

cor(mgr, method="spearman")

# produce one window with scatterplots of all columns against all other columns

pairs(mgr)

# you can see some clear correlations here!

# Which variables are they? Do they make sense?

#####################################

# 3. LINEAR REGRESSION

#####################################

fit <- lm ( GenusRich ~ SST.MEAN , data = mgr )

summary(fit)

# the plot command below on an 'lm' object will give you four

# diagnostic plots (for help: http://www.statmethods.net/stats/rdiagnostics.html)

par(mfcol=c(2,2)) # this sets up the window so you can create 4 plots in 1 window

plot(fit)

# are your residuals normally distributed?

# does normality improve if we use the square root of genus richness?

fit <- lm ( sqrtGenusRich ~ SST.MEAN , data = mgr )

summary(fit)

# the plot command below on an 'lm' object will give you four

# diagnostic plots (for help: http://www.statmethods.net/stats/rdiagnostics.html)

par(mfcol=c(2,2)) # this sets up the window so you can create 4 plots in 1 window

plot(fit)

# MULTIPLE REGRESSION

# include more than one predictor variable

# use "+" to add in an additional variable

fit2 <- lm ( sqrtGenusRich ~ SST.MEAN + PHOS, data = mgr )

summary(fit2)

plot(fit2)

#####################################

# 4. AUTOMATED MODEL SELECTION

#####################################

# R can use an automated method to find the best combination of variables

# based on Akaike's Information Criterion (AICc)

# you need to first install the package MuMIn using the code below

install.packages("MuMIn") # you only have to do this once per package

library(MuMIn) # do this every time you open the package in R

# create the model with variables you think make biological sense to include

# for example...

fit2 <- lm ( sqrtGenusRich ~ SST.MEAN + PHOS, data = mgr, na.action = "na.fail" )

# then simply use the function 'dredge' to try every possible model combination

fit.sel <- dredge(fit2)

# View the resulting object - you can see each row represents a different model

# with a particular combination of variables.

# The rows are ranked in order of AICc (lowest is the best)

fit.sel

############################################################

####### FOR YOUR REPORT 3 ######

####### START ######

############################################################

## Create a linear regression model to explain macroalgae genus richness.

## Find the best combination of variables using dredge.

## Use the code above to guide you.

## NOTE: this is the same exercise you did last week but on a different dataset!!!

## Think about hypotheses a priori (before you start!)

############################################################

####### FOR YOUR REPORT 3 ######

####### END ######

############################################################