

```

#=====
# file:          c:\Projects\CBP\Rcourse\FactorsRbit004.r
# function:      length weight regression for snook
#
# programmer:    Elgin S. Perry, Ph. D.
#
# date:         6/18/2014
#
# address:      2000 Kings Landing Rd.
#               Huntingtown, Md. 20639
#
# voice phone:  (410)535-2949
# email:        EPERRY@chesapeake.net
#=====

#install.packages()
#library(lattice) #Used for contour plots [contourplot()]
#library(nlme)    #used for gam Mixed model [gamm()]
#library(MASS)   #used for glm Mixed model [glmPQL()]
#library(mgcv)   #Wood's gam package
#library(chron)  #date functions
#library(doBy)   # Allows "BY processing similar to SAS
#library(FitAR)  #AR package from McLeod and Zhang
#library(Hmisc)  #stat function by Frank Harrell
#library(cluster) #cluster analysis routines
options(stringsAsFactors = FALSE)

# be sure to change \ to /
ProjRoot <- 'c:/Projects/CBP/Rcourse/'
setwd(ProjRoot);

# read data as per usual
datafile <- paste(ProjRoot,"snook.tdf",sep='');
snook <- read.table(datafile, header=TRUE, sep="\t", na.strings="NA", dec=".",
strip.white=TRUE,stringsAsFactors = FALSE)
# set erroneously recorded data to NA (missing)
snook[snook$length==40 & snook$water.body=='Atlantic' & snook$season=="May-Oct", 'wgt.mean'] <-
NA
#[1] "length"      "water.body"  "season"      "wgt.mean"   "wgt.min"    "wgt.max"

# use the factor function to create a new column in the snook data frame called
'water.body.f'
snook$water.body.f <- factor(snook$water.body)
boxplot(length~water.body.f,data=snook)

# use levels argument to control the order of levels
snook$water.body.f <- factor(snook$water.body,levels = c('Gulf','Atlantic'))
boxplot(length~water.body.f,data=snook)

# use lables argument to rename the levels
snook$water.body.f <- factor(snook$water.body,levels = c('Gulf','Atlantic'),labels=c('West
Coast', 'East Coast'))
boxplot(length~water.body.f,data=snook)

# other functions that can be used with factors
as.numeric(snook$water.body.f)
as.numeric(snook$water.body)
levels(snook$water.body.f)
levels(snook$water.body)
is.factor(snook$water.body.f)
is.factor(snook$water.body)
unclass(snook$water.body.f)
is.factor(unclass(snook$water.body.f))

```

```
# factors and linear models
lm1 <- lm(wgt.mean ~ water.body, data=snook)
summary(lm1)

lm2 <- lm(wgt.mean ~ water.body.f, data=snook)
summary(lm2)

# example of confusion
# first simulate some data
# create indices 1-120 to index the records of data
i <- 1:120
# create a vector x that follows half of a sign function
x <- sin(pi*i/120)
# add some noise to x to create y
y <- x + rnorm(120,0,1)
# create month indices by truncating the record indices
month.num <- floor((i-1)/10)+1
# create a vector of month names
months <-
c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December')
# create a vector of month name abbreviations
months.abr <- c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec')
# create a data frame using the vectors created above as columns
test <- data.frame(index = i, x=x, y=y, month.num=month.num)
# assign the character month strings to each row
test$month <- months[month.num]
test$month.abr <- months.abr[month.num]

# look at box plot by original numeric
boxplot(y~month.num, data=test)

# use factor function with default levels
test$month.f <- factor(test$month)
# now the months are displayed in alphabetical order not chronological order
boxplot(y~month.f, data=test)
# even more confusing is this plot using numeric values of the factor
boxplot(y~as.numeric(month.f), data=test)

#either of these two statements will assign levels to the months in chronological order
test$month.f <-
factor(test$month, levels=c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'))
#or
test$month.f <- factor(test$month, levels=months)

boxplot(y~month.f, data=test)
boxplot(y~as.numeric(month.f), data=test)
```