Lab 1

Rowan Edwards, Lois Deschaux, Ceili DeMarais

# Lab 1 Instructions:

Working in a group of 2, 3, or 4 people (randomly assigned in class, singletons allowed on this first lab too if you miss the in-person session), complete the following questions. Turn in a single PDF from your word document for the group *with all group member names on it* after knitting this document with your answers “in-line” (after the questions). Part of the assignment is to create a document that can successfully knit. You can discuss this lab with people outside of your lab group, but you must document any sources of information outside of the instructors, your group members, and the materials I am providing you (like web resources or other folks that you might ask for help).

Note that you have free access to Adobe Acrobat Pro via MSU at <https://www.montana.edu/uit/students/adobe/index.html> and will find that useful for merging separate files into a single PDF this semester.

Make sure to run spell-check using the “ABC check” button near the filename or use (i.e., pay attention to) the text underlining suggestions in the document (like for “filename”). The word document will not have spell-check turned on when built from a .Rmd file, so do the checking here.

The initial header contains some packages we will use frequently this semester. In addition to packages from CRAN, we need two special packages available from my github repository. In order to access them, you need to download and load the remotes package. Then uncomment the lines that contain #remotes::install\_github("greenwood-stat/ggResidpanel") and #remotes::install\_github("greenwood-stat/catstats2"). Do this to download the packages and then re-comment those lines as you will not need to do that again unless you update your version of R or change computers. Do NOT use install.packages("ggResidpanel") as this will install a different version. Then you can uncomment the #library(ggResidpanel) line for this lab and future uses.

## Factors related to intra-tendinous morphology of Achilles tendon in runners

To get a chance to review and practice/learn to use R, we will explore the data set posted to accompany Ho *et al*.’s “Factors related to intra-tendinous morphology of Achilles tendon in runners”. Please read the paper prior to January 17 to prepare for our first lab and we will revisit it for many more analyses.

* Ho K-Y, Baquet A, Chang Y-J, Chien L-C, Harty M, Bashford G, et al. (2019) Factors related to intra-tendinous morphology of Achilles tendon in runners. *PLoS ONE* 14(8): e0221183. <https://doi.org/10.1371/journal.pone.0221183>

It is available at <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221183>

* You will have time in the lab to work collaboratively to answer the following questions.

## General Questions:

1. What does PSFR stand for? Are higher or lower levels of it “good”?

* **PSFR stands for peak spatial frequency radius. PSFR was used to detect intra-tendinous morphological change, which was defined by collagen fiber disorganization. A low PSFR was associated with greater collagenous disarray, suggesting degeneration. Thus, higher levels of PSFR are “good”.**

1. What two predictor variables did they indicate were [“s”-word] related to the PSFR in their univariate models?

* **Male sex and presence of current Achilles tendon pain were indicated as [“s”-word] related to the PSFR in their univariate models.**

1. Which predictor did they indicate was [“s”-word] related to the PSFR in the “multivariate” model? What was the direction of the result (in other words, which levels/direction was related to “better” and “worse” on the response)?

* **Their multivariate model indicated that sex was [“s”-word] related with PSFR. Male runners displayed “worse” (lower) PSFR levels on average. \*\*Male and female are levels of the predictor variable.**

## Data Exploration:

The following code will get you started with the data set, including reading it in and wrangling the data set a bit.

* The data set is provided two possible ways. The authors’ posted it as an Excel spreadsheet. I also provided the data set in the catstats2 package under the name TendonData.
* The file “Runner Factors related to Intra-tendonious Morphology updated.xlsx” is posted on D2L. Download the .xlsx file and save it into a *new* folder, where you will store the data **and** today’s data analysis .Rmd file. You can choose the folder - maybe make a new one for our suite of the initial 3 to 5 labs that will use this same data set or make a new folder for each lab. Then make sure the readxl (for reading in xlsx files) package is installed (use the Packages tab or yellow bar at top of the .Rmd when you save it to tell you about a missing package used in the code below) and load it using library(readxl).
* The top of the .Rmd file has settings to eliminate reporting of messages and warnings (and significance stars!), but you will see code below to turn those on in certain codechunks options using message = T and warning = T. Having these off either in the codechunks or in the header above will clean up what appears in the final document but can make it more difficult to see mistakes in code sometimes.
* It is good code writing practice to keep a set of all related data, analysis, plots, documents, etc. in the same or possibly nested folders. When all the pieces are in the same folder, it allows for a clean workflow and working directory. When you are executing an R markdown file, R will search for things (such as data) in the **same** folder as where it is saved. If you do this, you will not need to specify anything other than read\_excel("filename.xlsx") in the code chunk.

*If* you are having troubles loading your data into RStudio and you have saved your files in this way, it is possible that R is searching in the wrong location and you need to change your working directory.

* The easiest way to do this is,
  + click on the **Session** drop-down from the top of the screen,
  + select the **Set Working Directory** tab,
  + select **To Source File Location**.
  + Copy the code into the code chunk before the code to read in the data set.
* After this process R will be searching for objects (such as data) in the **same** folder that the document/script you’re working on is saved in.

Now we can read the data set into our R session using read\_excel("file.xlsx").

## General information on reading in files in RStudio:

* Note that if you are used to using the “Import Dataset” button in RStudio, you can select whether to use the read\_csv function from readr or read.csv; read\_csv reads the data set in as a tibble and will be our most typical format. If you use read.csv, it will be read the data in as a data.frame. You can use either. The read\_... functions make fewer assumptions about the variable types than read.csv and you will need to transform any categorical variables as factors (using factor) - see the mutate() part of the code below that is done for you this one time. Text variables are read in as character type variables in the read\_... functions and it is our task to define them as factors (and to define any numerically coded variables that represent factor levels as factors). There are options to modify (most easily using the “Import Dataset” button) if your missing data are stored as anything other than blank cells or the data are in any other format (SPSS, SAS, or STATA data formats).
* The readxl package and its read\_excel function can read data sets directly from Excel spreadsheets. This allows you to maintain multi-tab excel data sets, read in from the desired page, and even select the rows to read in. The potential for keeping multiple sheets together (especially if the other sheets contain metadata (source of information and variable definitions, for example) or different related data sets) make the choice of Excel files more attractive than they were 10 years ago. But .csv files are still considered more stable ways of archiving data for future use, but they only have a single “sheet”.

## Another option for accessing the data:

* If you have problems reading in the data set, you can try to load it directly from the catstats2 package to have it available using data(TendonData). It also has a help file that contains more information about the data set and some possible code. After the first lab, we will switch to using the version of the data set from catstats2.

## Data wrangling

* It is a good idea to explore the data set briefly. Explore the results of View(tendon), head(tendon), and tail(tendon) by uncommenting each line (you can re-comment them to submit the lab). Or to look more generally using a function like glimpse that is already un-commented.

library(readxl)  
 tendon <- read\_excel("Runner Factors related to Intra-tendonious Morphology updated .xlsx")  
  
 #If using data(TendonData) from catstats2:  
 #tendon <- TendonData  
  
 #View(tendon)  
 #head(tendon)  
 #tail(tendon)  
  
 glimpse(tendon)

## Rows: 182  
## Columns: 15  
## $ `Subject ID` <chr> "UNLV001", "UNLV002", "UNLV003", …  
## $ Side <chr> "L", "L", "L", "L", "L", "L", "L"…  
## $ PSFR <dbl> 2.115342, 1.998679, 2.032770, 1.9…  
## $ HeelRaise <dbl> 31, 21, 30, 41, 20, 17, 21, 17, 1…  
## $ KneeWall <dbl> 13.0, 21.0, 13.0, 15.0, 14.0, 13.…  
## $ `CurrPain (Y=1)` <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, …  
## $ `Neovascularization (Doppler) (Y=1)` <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ `VISA-A` <dbl> 98, 100, 100, 90, 96, 95, 99, 100…  
## $ Age <dbl> 31, 30, 33, 30, 45, 37, 44, 30, 2…  
## $ `Sex (M=1)` <dbl> 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, …  
## $ YearsRunning <dbl> 4, 17, 10, 4, 7, 10, 15, 2, 4, 15…  
## $ BMI <dbl> 20.04244, 18.87405, 17.84817, 22.…  
## $ `Waist to Hip` <dbl> 0.7446809, 0.7127660, 0.7272727, …  
## $ CSA <dbl> 54.298, 57.771, 57.310, 51.554, 4…  
## $ `Hx Pain (Y=1)` <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, …

* The following code leverages some functions from the tidyverse to change names of some unfortunately named variables (rename), the mutate\_if(is.character, as.factor) is a neat bit of code that looks for character variables and turns them into factor variables (here that is only two variables, SubjectID and Side), and the mutate function is used to change three numerically coded factor variables into factors so they will be properly modeled by R and create two new variables related to the location and combinations observed for current and past pain responses. Some of these variables we will explore on later labs.

library(tidyverse)  
  
 tendon <- tendon %>%  
 dplyr::rename(SubjectID = 'Subject ID',  
 Sex = 'Sex (M=1)',  
 CurrPain = 'CurrPain (Y=1)',  
 WaisttoHip = 'Waist to Hip',  
 VISAA = 'VISA-A',  
 Neovascularization = 'Neovascularization (Doppler) (Y=1)',  
 HistPain = 'Hx Pain (Y=1)') %>%  
 mutate\_if(is.character, as.factor) %>%  
 mutate(CurrPain = factor(CurrPain),  
 Neovascularization = factor(Neovascularization),  
 Sex = factor(Sex),  
 Location = factor(substr(SubjectID, 1, 3)),  
 HistPain = factor(HistPain),  
 PainCombs = factor(str\_c(HistPain, CurrPain)))  
 levels(tendon$PainCombs) <- c("NoPain", "PreviousPainOnly", "CurrentPainOnly", "Both")

1. They mention using a linear mixed model but their explanation is not completely clear. One assumption of using a regular linear model (our lm) is that all the observations taken are independent. Discuss how the data they collected violate the independence assumption if a linear model had been used for modeling PSFR. Just base this on their data collection explanation and who/what they measured and modeled, reviewing the data set help with understanding what they measured on each subject.

* **There could have been multiple clustering effects in their study design related to multiple samples taken from the same individual, two different locations, sex, age, lifestyle, and other related factors. The use of a regular linear model in this case could result in inaccuracies because of possible correlations within subject measurements. They could instead fit a mixed linear model that would account for these dependencies.**

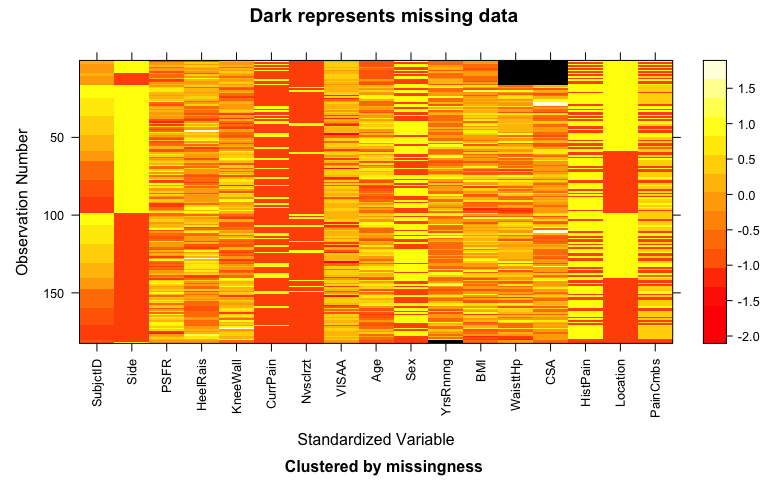
1. The following code provides a plot that can help us to visualize responses and, more importantly, missing data in our data set. Which variables contain missing data and which ones have the most missing observations (you do not need to find the exact numbers)?

* **Years running, waist to hip ratio, and CSA variables all have missing observations. It seems as though waist to hip and CSA contain the most missing observations.**

library(mi)  
tdf <- missing\_data.frame(data.frame(tendon))

## NOTE: The following pairs of variables appear to have the same missingness pattern.  
## Please verify whether they are in fact logically distinct variables.  
## [,1] [,2]   
## [1,] "WaisttoHip" "CSA"

image(tdf)



table(tdf@patterns)

##   
## nothing YearsRunning WaisttoHip, CSA   
## 164 2 16

1. The following code creates two different versions the data set. What happened in creating tendonna1 and tendonna2 and what are the sample sizes?
   * **‘tendonna1’ removed the samples what were missing the observations for CSA, bringing the overall sample size down to 166. ‘tendonna2’ removed samples with the remaining NA missing observations, leaving a sample size of 164.**

tendonna1 <- tendon %>% drop\_na(CSA)  
tendonna2 <- tendon %>% drop\_na()  
dim(tendon)

## [1] 182 17

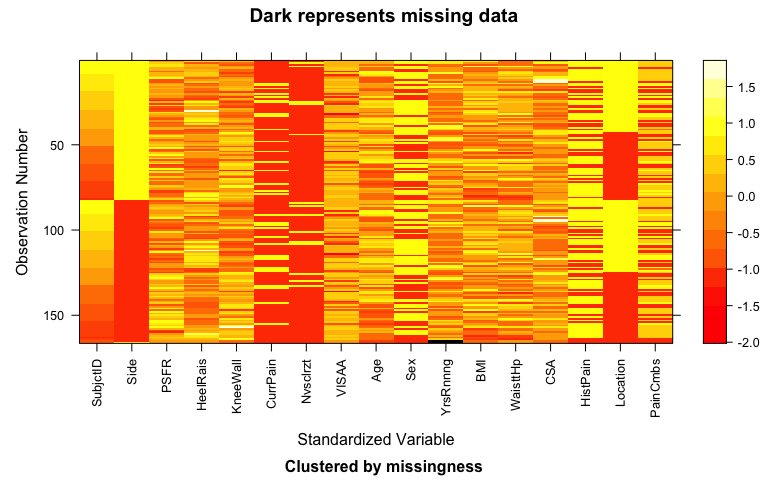
dim(tendonna1)

## [1] 166 17

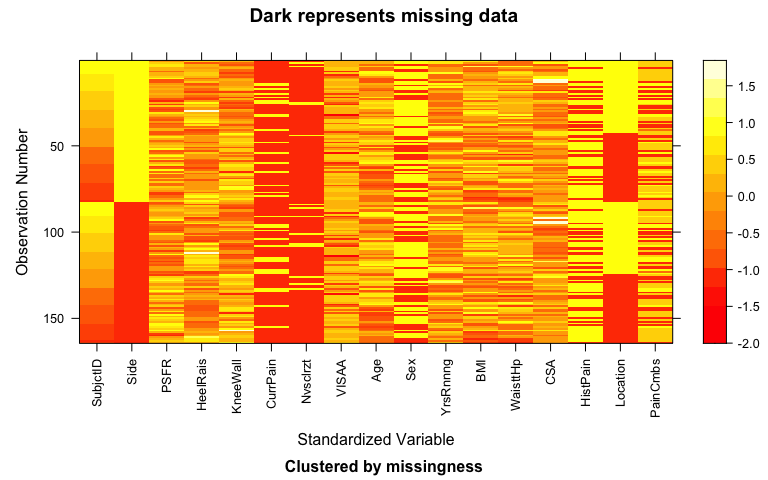
dim(tendonna2)

## [1] 164 17

tdf1 <- missing\_data.frame(data.frame(tendonna1))  
image(tdf1)

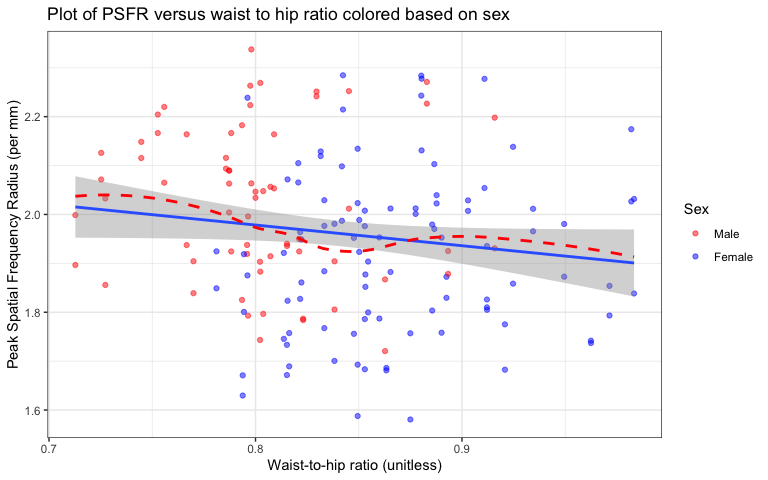


tdf2 <- missing\_data.frame(data.frame(tendonna2))  
image(tdf2)



1. We will use tendonna2 moving forward (the researchers used “complete case” versions of each model so the univariate and “multivariate” model results had differing sample sizes, as we will explore later). Use ggplot to make a plot of PSFR (y-axis) vs Waist-to-Hip ratio (x-axis, WaisttoHip) coloring the points by Sex of subjects and adding lm and nonparametric smoothing lines (hint: you need to add geom\_point and two geom\_smooth’s). Also modify the title of the plot. To help you, the code is started for you:

library(ggplot2)  
tendonna2 %>% ggplot(mapping = aes(x = WaisttoHip, y = PSFR)) +  
 geom\_point(mapping = aes(col = Sex), alpha = 0.5) +  
 geom\_smooth(method = "lm") +  
 geom\_smooth(col = "red", lty = 2, se = F) +  
 scale\_color\_manual(values = c("0" = "red", "1" = "blue"),  
 labels = c("Male", "Female")) +  
 labs(x = "Waist-to-hip ratio (unitless)",  
 y= "Peak Spatial Frequency Radius (per mm)",  
 title = "Plot of PSFR versus waist to hip ratio colored based on sex")



1. I like the general structure of the paper and, after a few emails and them updating the posted data set, we can more or less replicate their results. We’ll eventually incorporate the mixed model ideas similar to what they used, but for now use tendonna2 to fit an lm with PSFR as the response and Sex and WaisttoHip as the explanatory variables (no interaction). Report a model summary() and also run something like modelname %>% tbl\_regression(intercept = T) for your model (no discussion, just fit the model and include the output).

model <- lm(PSFR ~ Sex + WaisttoHip, data = tendonna2)  
summary(model)

##   
## Call:  
## lm(formula = PSFR ~ Sex + WaisttoHip, data = tendonna2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33903 -0.11799 0.00947 0.10616 0.36849   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.93675 0.21261 9.110 3.12e-16  
## Sex1 -0.11183 0.03134 -3.568 0.000475  
## WaisttoHip 0.10826 0.26421 0.410 0.682535  
##   
## Residual standard error: 0.1625 on 161 degrees of freedom  
## Multiple R-squared: 0.09293, Adjusted R-squared: 0.08166   
## F-statistic: 8.247 on 2 and 161 DF, p-value: 0.0003892

model %>% tbl\_regression(intercept = T)

| **Characteristic** | **Beta** | **95% CI***1* | **p-value** |
| --- | --- | --- | --- |
| (Intercept) | 1.9 | 1.5, 2.4 | <0.001 |
| Sex |  |  |  |
| 0 | — | — |  |
| 1 | -0.11 | -0.17, -0.05 | <0.001 |
| WaisttoHip | 0.11 | -0.41, 0.63 | 0.7 |
| *1*CI = Confidence Interval | | | |

1. It’s useful to record some information about the version of R you are using. When you Knit this documentation, it will report on the version of R that you are using. It should say 4.4.2 in your compiled (knitted) word document:

* R version (short form): 4.4.2

1. Document any resources used outside of your fellow group members and course provided resources. If you do not use any, report “NONE” to get credit for this question.

* NONE