Lab 1

Hunter Charles, Macy Collins, Jared Adam

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. Do not knit directly to PDF in this class - the code templates are not designed to do this correctly.

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.

White matter lesions:

To get a chance to review and practice/learn to use R, we will explore the data set posted to accompany Biner et al.'s (2015) "Factors related to intra-tendinous morphology of Achilles tendon in runners". Please read the paper prior to August 23 to prepare for our first lab and we will revisit it for many more analyses.

- Birner A, Seiler S, Lackner N, Bengesser SA, Queissner R, Fellendorf FT, et al. (2015) Cerebral White Matter Lesions and Affective Episodes Correlate in Male Individuals with Bipolar Disorder. PLoS ONE 10(8): e0135313. doi:10.1371/journal.pone.0135313

- Use the link to download the paper or use the provided pdf of the paper on D2L. We will get further into what they did a bit later on.
- The data set is provided on D2L as a .csv file. Download the file and save it into a *new* folder, where you will store the data **and** today's data analysis file.
- It is good code writing practice to keep a set of all related data, analysis, plots, documents, etc. in the same folder. When all the pieces are in the same folder, it allows for a clean workflow and working directory. When you are executing code for a document/script `R` will search for things (such as data) in the **same** folder as the document/script. 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 simply read the data set into R using `read_csv("file.csv")` (this does require the `readr` (Wickham et al., 2024) R package (R Core Team, 2024))
- An example of a couple of citations:
  - Wickham H, Hester J, Bryan J (2024). *readr: Read Rectangular Text Data*. R package version 2.1.5
  - R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

```
## 1. Make sure this RMarkdown file is saved in the SAME folder as the
bipolar data
## 2. Now load the bipolar data into R

bpdata <- read_csv("bipolardata.csv")

#View(bpdata)

#head(bpdata)

#tail(bpdata)
```

```
glimpse(bpdata)
```

```
## Rows: 154
## Columns: 28
## $ ID                      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
14, …
## $ age1                    <dbl> 29, 29, 49, 66, 21, 36, 25, 28, 39, 63, 54,
41,…
## $ `Illness duration`      <dbl> 18, 7, 5, 9, 6, 9, 2, 12, 31, 48, 35, 16,
34, 4…
## $ nWML_sum                <dbl> 0.1666159, 0.2236509, 0.2469235, 0.2579945,
0.2…
## $ WML_volume              <dbl> 2439.351, 3888.340, 3775.342, 4038.991,
3731.03…
## $ depression              <dbl> 4, 5, 10, 6, 3, 6, 3, 10, 12, 10, 10, 0, 6,
13,…
## $ mania                   <dbl> 2, 5, 9, 6, 4, 6, 2, 2, 6, 20, 10, 1, 10,
4, 5,…
## $ sex                     <dbl> 1, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2,
2, 2,…
## $ group                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
## $ LITH_AKT                <dbl> 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ ATYPS                   <dbl> 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1,
0, 1,…
## $ ANTIE                   <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
1, 1,…
## $ AD                      <dbl> 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1,
1, 0,…
## $ bip1_bip2               <dbl> 2, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1,
2, 1,…
## $ BMI                     <dbl> 27.83763, 30.93044, 33.88175, 26.54321,
23.9614…
## $ MIGR                    <dbl> 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ BEG_PSY_S_age           <dbl> 13, 23, 44, 59, 14, 26, 23, 16, 8, 15, 19,
26, …
## $ BDI                     <dbl> 9, 0, 21, 0, 11, 12, 15, 8, 10, 6, 3, 1, 8,
0, …
## $ HAMD                    <dbl> 4, 0, 13, 0, 6, 12, 5, 0, 9, 12, 3, 0, 0,
2, 7,…
## $ YMRS                    <dbl> 1, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, NA…
## $ smoking_yes_no          <dbl> 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0,
0, 1,…
## $ LnWML_sum               <dbl> 0.1541072, 0.2018389, 0.2206793, 0.2295188,
0.1…
## $ RES_LnWMLsum            <dbl> -0.033701675, 0.045323544, 0.055769031, -
```

```
0.1004…
## $ HYPERT                  <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1,
1, 0,…
## $ DM                       <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ `substance dependancy` <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1,…
## $ `alcohol dependancy`   <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0,…
## $ `anxiety disorder`     <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1,…
```

- If you are used to using the "Import Dataset" button in RStudio, you can select whether to use the `read_csv` function from `readr` instead of `read.csv`, which will read the data set in as a `tibble`. If you use `read.csv`, it will be read the data in as a `data.frame`. You can use either, with tibbles you will need to transform any categorical variables as factors (using `factor` within a `mutate` string (see example below)) to use the functions we use. In `data.frames`, assumptions are made about what is factor and what is numeric or character strings and can then require processing after the data are read in to deal with some ways variables are coded, especially when missing data are present and coded in some way other than empty cells. This course focuses on using `tidyverse` methods.

- Another option in RStudio is to use the `readxl` package and its `read_excel` function to read data sets directly from Excel spreadsheets (this function figures out whether it is a .xls or .xlsx file). Storing data in Excel allows you to maintain multi-tab Excel data sets, read in from the desired page, and even select the rows to read in. I posted the `datafile_bipolar.xlsx` file and it contains the same information as the .csv file in the `SourceDataWMLBirner` tab. It also contains some "meta-data" about the variables in the `Variables` tab. You will need that information to answer some questions. I would not read that into R but the utility of working with Excel spreadsheets is illustrated in this situation to be able to store "metadata" in the same file but in a way that doesn't contaminate the data itself.

- It is a good idea to explore the data set briefly. Explore the results of `View(bpdata)` (you will need to uncomment), `head(bpdata)`, and `tail(bpdata)`.

- We will focus on the `nWML_sum` as a response variable (renamed below to `NormVolume`). This is their "volume of cerebral white matter lesions" measurement in cubic millimeters divided by total intracranial volume. For some reason it can exceed one. They call this `normalization` although that term more typically means re-scaling a response to have mean 0 and SD of 1 (a z-score). This is a ratio response that is like a rate (volume of lesions over total volume).

- The following code leverages some functions from the `tidyverse` to change names of some unfortunately named variables (`rename`) and the `mutate` function is used to change some of the numerically coded factor variables into factors and the

`fct_recode` function is used to incorporate some of the metadata on the variables into more explicit factor labels.

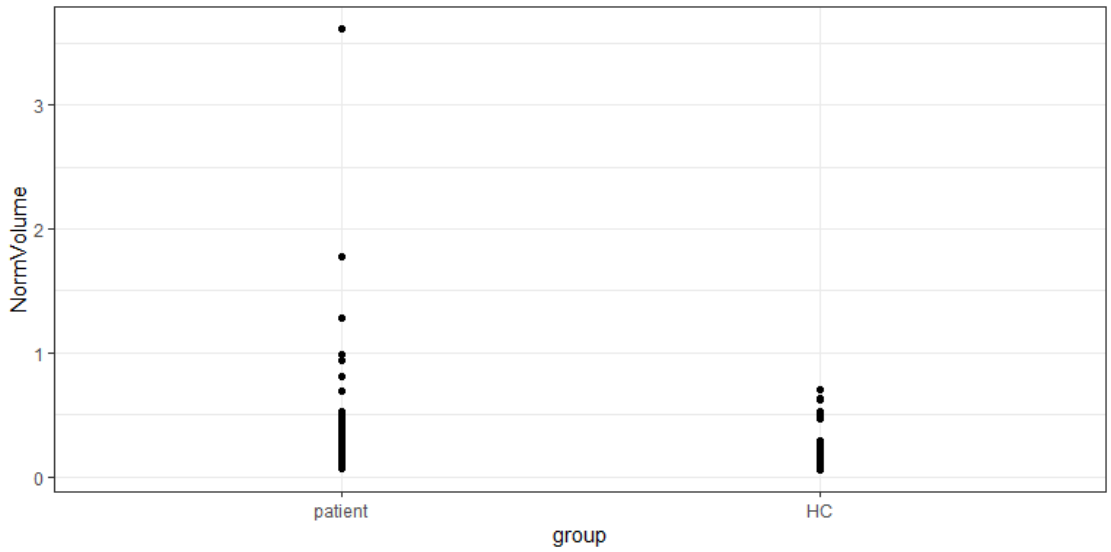
```
 library(tidyverse)

bpdata <- bpdata %>% dplyr::rename(IllnessDuration = 'Illness duration',
                                   NormVolume = 'nWML_sum',
                                   SubDep = 'substance dependancy',
                                   AlcDep = 'alcohol dependancy',
                                   AnxDis = 'anxiety disorder',
                                   smoking = 'smoking_yes_no') %>%
      mutate(group = factor(group),
             sex = factor(sex),
             YMRS = factor(YMRS),
             DM = factor(DM),
             smoking = factor(smoking),
             HYPERT = factor(HYPERT),
             group = fct_recode(group,
                                patient = "1",
                                HC = "2"),
             sex = fct_recode(sex,
                              male = "1",
                              female = "2"),
             smoking = fct_recode(smoking,
                                  no = "0",
                                  yes = "1"),
             GroupSexCombs = factor(str_c(group, sex)) #For later use
             )
```

1) Generate and report the summary statistics using the `favstats` function from `mosaic` for the `NormVolume` variable both overall and for `BD` and `HC` groups. Report your code and output. No discussion.

```
library(mosaic)
favstats(NormVolume ~ 1, data = bpdata)

##   1       min        Q1   median       Q3      max      mean        sd   n
## 1 1 0.0561514 0.1373177 0.211394 0.311464 3.615256 0.2862283 0.3473254 154
##   missing
## 1       0

favstats(NormVolume ~ group, data = bpdata)

##     group        min        Q1    median        Q3       max      mean
## 1 patient 0.06890985 0.1727848 0.2394119 0.3435528 3.6152561 0.3359443
## 2      HC 0.05615140 0.1127530 0.1405374 0.2071285 0.7028555 0.1941617
##          sd   n missing
## 1 0.4092558 100       0
## 2 0.1486240  54       0
```

2) Plot the `NormVolume` variable by `group`. Report your `R` code and plots.

```r
ggplot(bpdata, aes(group, NormVolume))+
  geom_point()
```



3) Generate a potentially appropriate statistical test to compare the two groups. Report R code and output. No discussion.

```r
lm1 <- lm(NormVolume ~ group, data = bpdata)
summary(lm1)

##
## Call:
## lm(formula = NormVolume ~ group, data = bpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2670 -0.1249 -0.0730  0.0112  3.2793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33594    0.03417   9.830   <2e-16
## groupHC     -0.14178    0.05771  -2.457   0.0151
##
## Residual standard error: 0.3417 on 152 degrees of freedom
## Multiple R-squared:  0.03819,    Adjusted R-squared:  0.03186
## F-statistic: 6.035 on 1 and 152 DF,  p-value: 0.01515
```

4) In Table 1, they note that they used a model for the normalized WML volumes that include the BD/HC variable as its focus but "controlled for age, diabetes, smoking and BMI". Uncomment and modify the following code to improve the coding of the DM variable.

```r
bpdata <- bpdata %>% mutate(
  DM = fct_recode(DM,
                  no = "0",
                  yes = "1"))
```

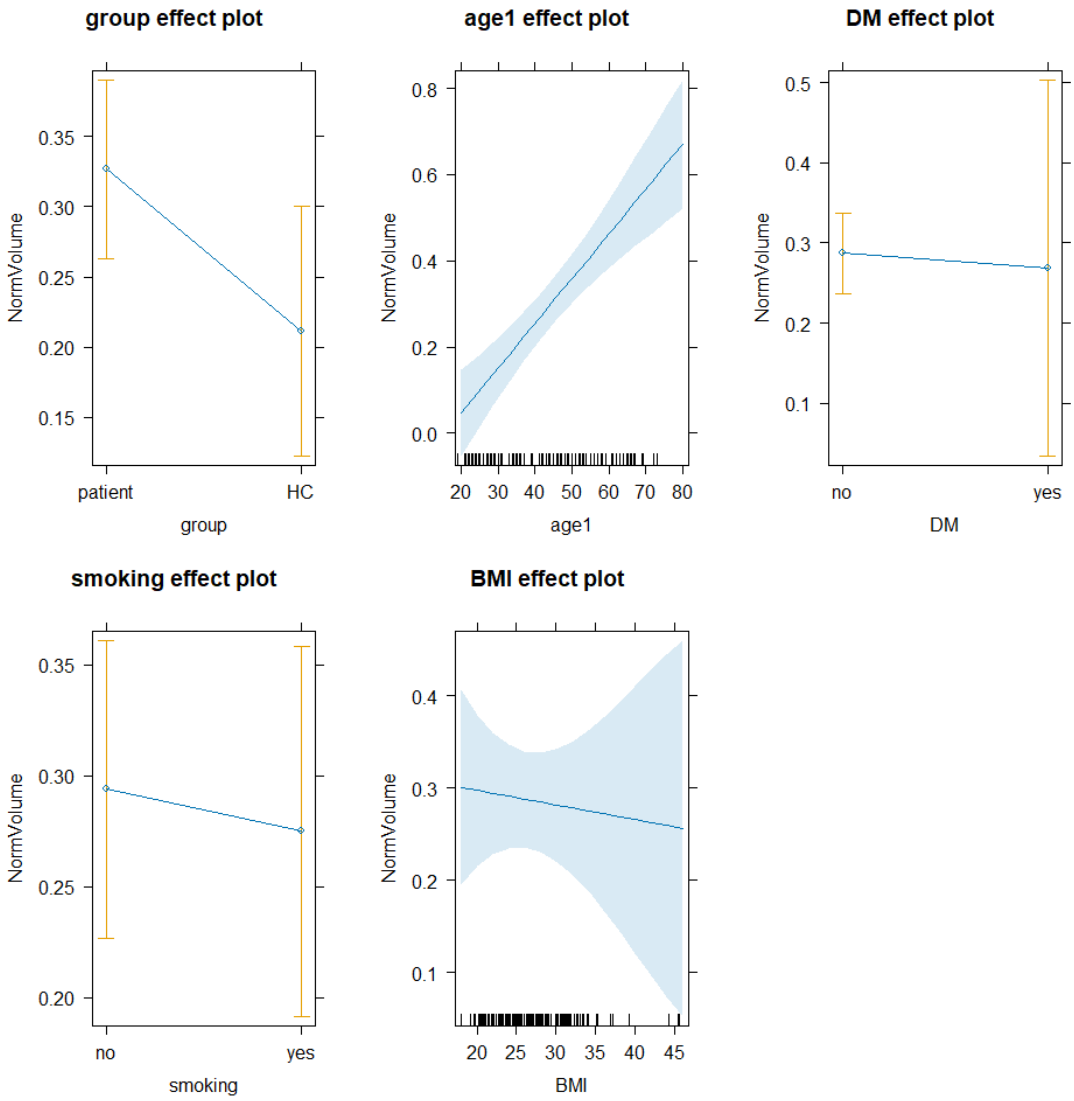
5) Fit the linear model they describe using `NormVolume` as the response variable using `group`, `age1`, `DM`, `smoking` and `BMI` as explanatory variables (no interactions). Report your `R` code and a model `summary()`.

```
lm2 <- lm(NormVolume ~ group + age1 + DM + smoking + BMI, data = bpdata)
summary(lm2)
```

```
##
## Call:
## lm(formula = NormVolume ~ group + age1 + DM + smoking + BMI,
##     data = bpdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38141 -0.13509 -0.03054  0.07240  2.88046
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.068098   0.164599  -0.414   0.6797
## groupHC     -0.115474   0.057677  -2.002   0.0471
## age1         0.010375   0.001920   5.405 2.54e-07
## DMyes       -0.018445   0.121506  -0.152   0.8795
## smokingyes  -0.018928   0.056867  -0.333   0.7397
## BMI         -0.001594   0.005253  -0.303   0.7620
##
## Residual standard error: 0.3097 on 148 degrees of freedom
## Multiple R-squared:  0.2307, Adjusted R-squared:  0.2047
## F-statistic: 8.876 on 5 and 148 DF,  p-value: 2.168e-07
```

6) Make and report an `effects` plot of the your model from the previous question. No discussion, just the plot.

```
# Install and load the effects package
library(effects)

plot(allEffects(lm2))
```



7) 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 4.4.1 in your compiled word document:

- R version (short form): 4.4.1