Lab 3

Samanthi Wijerathna, Rubey Berkenkamp

Lab 3 Instructions:

Working in a group of 2, 3, or 4 people, 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).

White matter lesions (continued):

```
bpdata <- read_csv("bipolardata.csv")

 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"),
             IllnessDurBin =
forcats::fct_explicit_na(cut_number(IllnessDuration, n = 2))
             )

#?fct_explicit_na # gives missing values a factor level
#?cut_number # balances observations between these two groups

favstats(IllnessDuration ~ group, data = bpdata)
```

```
##     group min Q1 median   Q3 max  mean       sd   n missing
## 1 patient   1  9     16 27.5  60 19.06 13.18847 100       0
## 2      HC  NA NA     NA   NA  NA   NaN       NA   0      54

favstats(IllnessDuration ~ IllnessDurBin, data = bpdata)

##   IllnessDurBin min Q1 median    Q3 max      mean        sd  n missing
## 1        [1,16]   1  5      9 13.00  16  9.096154  4.827453 52       0
## 2       (16,60]  17 20     29 35.25  60 29.854167 10.595040 48       0
## 3     (Missing)  NA NA     NA    NA  NA       NaN        NA  0      54
```

1) Illness duration is a complicated variable because it is always missing for healthy subjects (see `favstats` result above), which is a systematic or structural missingness pattern due to the definition of the variable. The previous code creates a new variable called `IllnessDurBin`. Explain the levels of the new variable and discuss how many observations are in each level.

The new variables has split the 100 patient observations into two bins that are roughly balanced in number of observations. The above code applies missing values as a factor level and then balances the illness duration into n = 2 bins, thus roughly balancing the illness groups with the missing/ healthy group.

2) While those levels are fairly explicit, they will be awkward to discuss and use. Use `fct_recode` to improve the names of the levels of `IllnessDurBin` with text labels for each level related to what each category is. When we rename levels, the levels now might not follow the alphabetical convention in R. What is the order of your new levels and does this differ from the typical alphabetical convention that R uses?

It is numerical, but we assigned new names.

```
levels(bpdata$IllnessDurBin)

## [1] "[1,16]"    "(16,60]"   "(Missing)"

?fct_recode

bpdata <- bpdata %>%
  dplyr::mutate(IllnessDurBin = fct_recode(IllnessDurBin,
                            short = "[1,16]",
                           long = "(16,60]",
                           healthy = "(Missing)"))
levels(bpdata$IllnessDurBin)

## [1] "short"   "long"    "healthy"
```

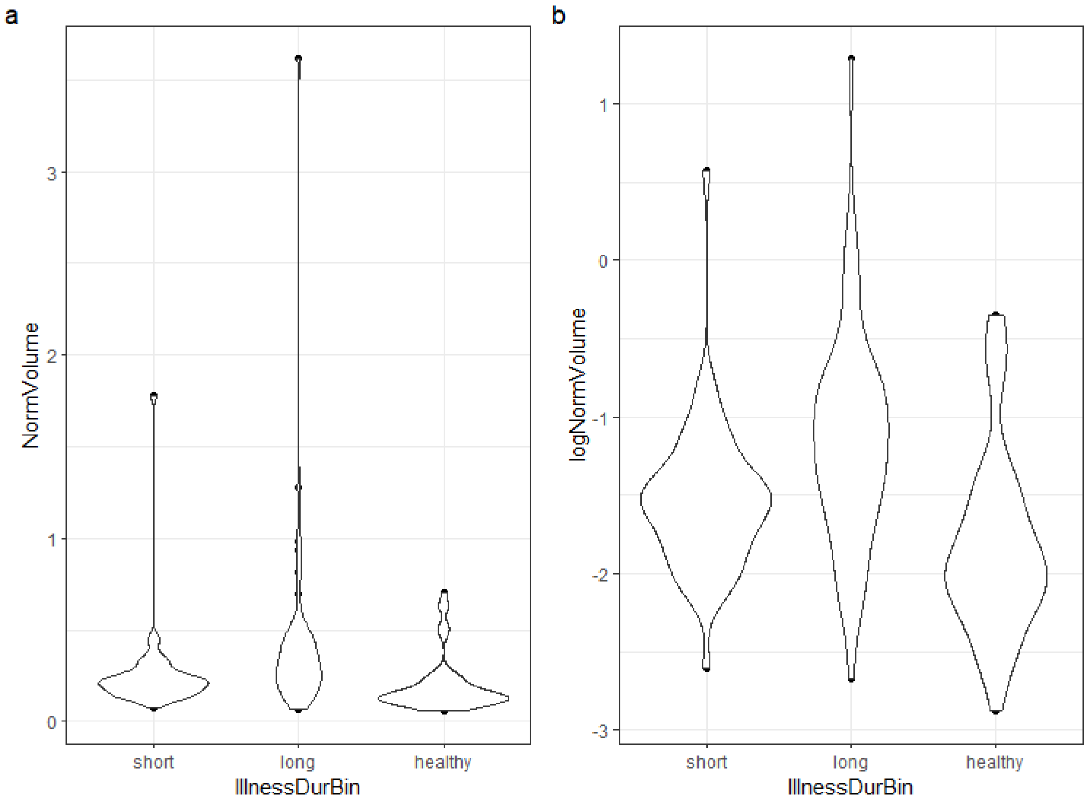
3) Eventually we will go through our diagnostics on the models with `NormVolume`, but it seems like we might need to consider the same transformation the author's used (log-transforming it). Why is the `favstats` as used below important to do prior to log-transforming a variable?

We need to understand the distribution of the raw data. Looking at numerical spread of a log transformed variable is too hard to interpret.

```
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

bpdata <- bpdata %>% mutate(logNormVolume = log(NormVolume))
```

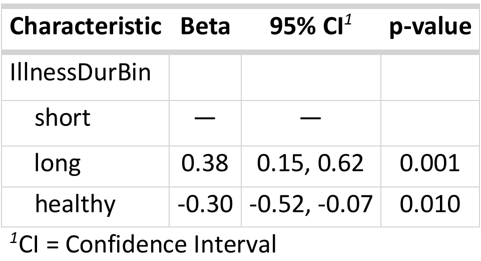
4) Make plots of the `NormVolume` and `logNormVolume` versus the `IllnessDurBin` and “patch” the two plots together using the `patchwork` package with tagging using `plot_annotation(tag_levels = "a")`. No discussion and you don’t need to improve the axis labels or title this time.

```
p1 <- ggplot(bpdata, aes(IllnessDurBin, NormVolume))+
  geom_point()+
  geom_violin()


p2 <- ggplot(bpdata, aes(IllnessDurBin ,logNormVolume))+
  geom_point()+
  geom_violin()


p1 + p2 + plot_annotation(tag_levels = 'a')
```



5) Fit a linear model for `logNormVolume` with `IllnessDurBin` as the only predictor, generate model summaries with `summary` and `tbl_regression` (making sure the intercept is included) and an effects plot. No discussion.

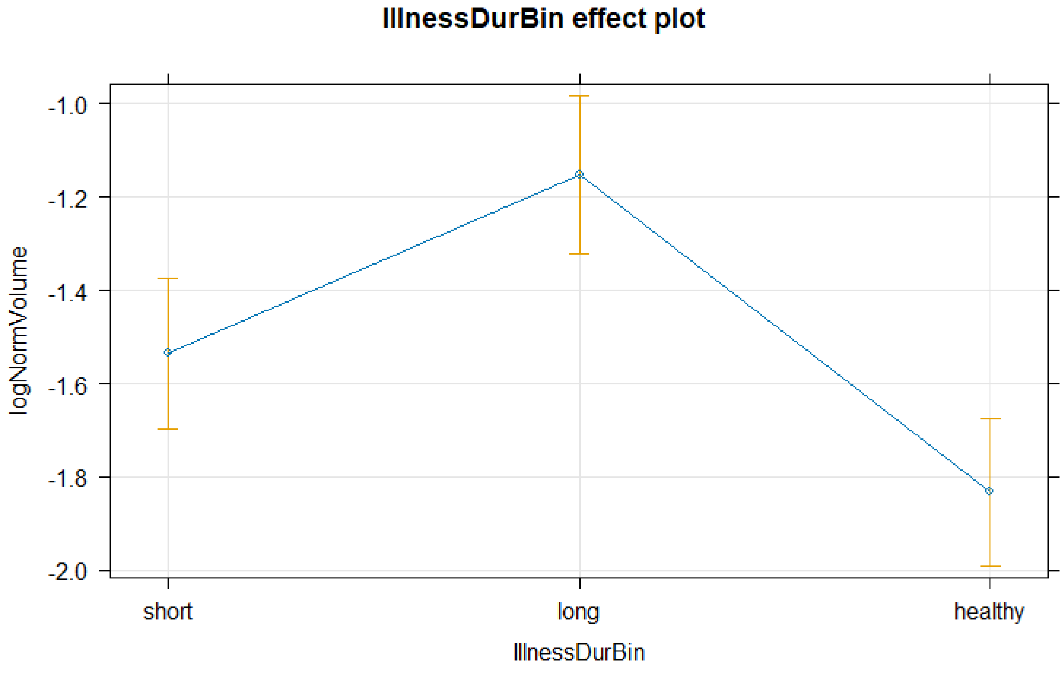
```
m1 <- lm(logNormVolume ~ IllnessDurBin, data = bpdata)
summary(m1)

## 
## Call:
## lm(formula = logNormVolume ~ IllnessDurBin, data = bpdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.52247 -0.34435 -0.05762  0.28892  2.43765 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -1.53504    0.08161 -18.809  < 2e-16
## IllnessDurBinlong     0.38255    0.11779   3.248  0.00143
## IllnessDurBinhealthy -0.29681    0.11434  -2.596  0.01037
## 
## Residual standard error: 0.5885 on 151 degrees of freedom
## Multiple R-squared:  0.1834, Adjusted R-squared:  0.1726 
## F-statistic: 16.96 on 2 and 151 DF,  p-value: 2.271e-07

tbl_regression(m1)
```



Characteristic	Beta	95% CI[1]	p-value
IllnessDurBin			
short	—	—	
long	0.38	0.15, 0.62	0.001
healthy	-0.30	-0.52, -0.07	0.010

[1] CI = Confidence Interval

```
plot(allEffects(m1), grid = T)
```



6) Write out the estimated model, defining the indicator variables you used.

$$\hat{\mu}\{logNormVolume|IllnessDurBin\} = -.153 + 0.38I_{Illness=Long} - 0.29I_{Illness=Healthy}$$

- where

$$I_{Illness=Long}$$

when a subject had a long illness duration is 1 and 0 otherwise.

- where

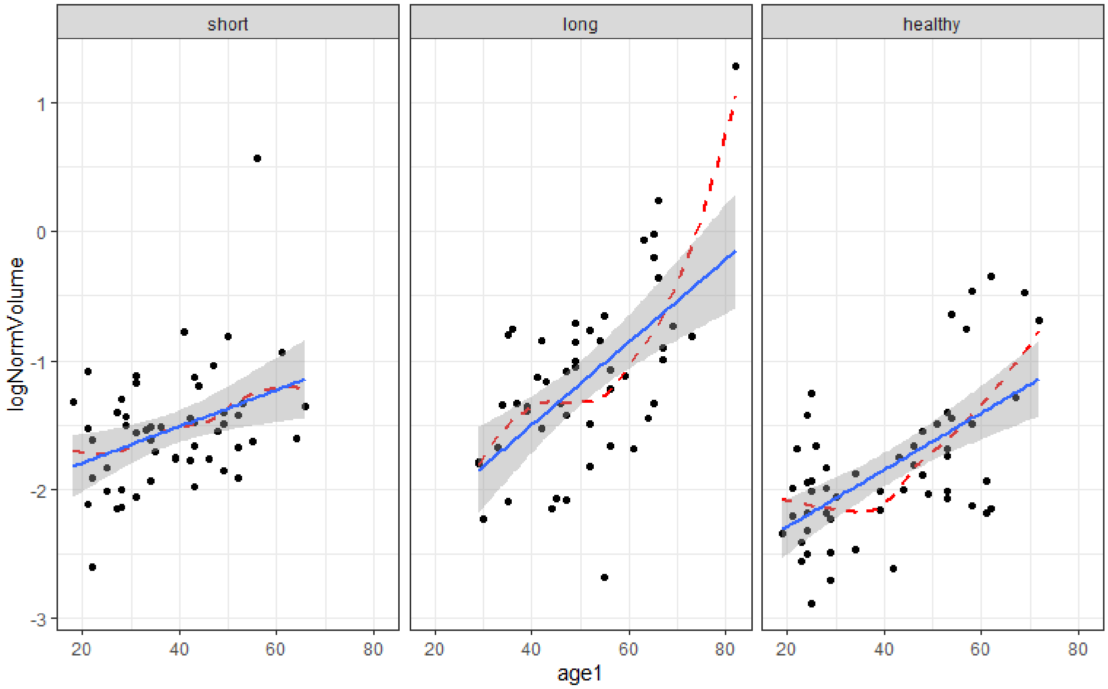
$$I_{Ilness=Healthy}$$

when a subject was healthy is 1 and 0 otherwise.

7) Make a plot of the `logNormVolume` based on `age1` faceted by `IllnessDurBin`. Add both linear and nonparametric smoothing lines (this hint will be removed into the future with this sort of plot and is just an assumption that you try to add these lines

and only remove if you have some reason). Discuss the potential for and pattern of an interaction based on the plot.

```
ggplot(bpdata, aes(age1, logNormVolume))+
  geom_point()+
  facet_wrap(~IllnessDurBin)+
  geom_smooth(se = F, lty = 2, col = 'red')+
  geom_smooth(method = 'lm')
```



There is a potential for an interaction because the slopes for logNormVolume are different among illness duration facets based on age group.

8) Fit the interaction model that relates to the previous plot, a model `summary`, and generate an effects plot. Report the estimated intercept and `age1` slope for a healthy subject (no illness) and one in the lower illness duration group, showing your work for the calculations. You do not need to write out the estimated model (for time purposes).

- Full model :

$$\hat{\mu}\{logNormVolume|IllnessDurBin * Age\} = -2.07 - 0.71I_{Illness=Long} - 0.66I_{Illness=Healthy} + 0.014Age + 0.018I_{Illness=Long} * Age + 0.008I_{Illness=Healthy} * Age$$

- Lower Illness duration group:

$$\hat{\mu}\{logNormVolume|IllnessDurBin$$
$$= Short * Age\} - 2.07 - 0.71 I_{Illness=Long} * 0 - 0.66 I_{Illness=Healthy} * 0 + 0.014Age$$
$$+ 0.018 I_{Illness=Long} * 0 * Age + 0.008 I_{Illness=Healthy} * 0 * Age$$

$$= -2.07 + 0.014Age$$

- Equation for healthy group:

$$= -2.07 - 0.71 * 0 - 0.66 I_{Illness=Healthy} * 1 + 0.014Age + 0.008 I_{Illness=Healthy} * 1 * Age$$

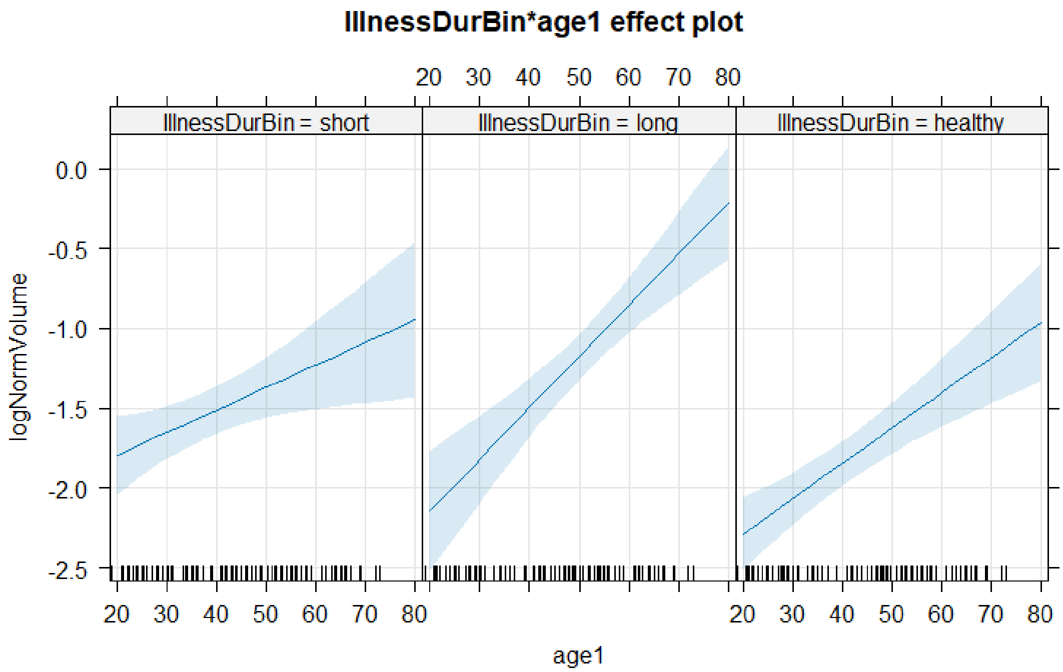
$$= (-2.07 - 0.66) + (0.014 + 0.008)$$

Combining like terms of intercept estimate and healthy estimate and combining like terms of turned on age components (age and age of healthy).

$$= -2.73 + 0.022$$

- Estimated intercept for lower illness duration group: -2.07
- Estimated age slope for lower illness duration group: 0.014
- Estimated intercept for healthy group: -2.73
- Estimated age slope for healthy group: 0.022

```
m2 <- lm(logNormVolume ~ IllnessDurBin*age1, data = bpdata)
summary(m2)

## 
## Call:
## lm(formula = logNormVolume ~ IllnessDurBin * age1, data = bpdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.65899 -0.33220 -0.01564  0.27694  1.85992 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.074873   0.227808  -9.108 5.39e-16
## IllnessDurBinlong        -0.707779   0.374645  -1.889   0.0608
## IllnessDurBinhealthy     -0.655589   0.297669  -2.202   0.0292
## age1                      0.014099   0.005670   2.487   0.0140
## IllnessDurBinlong:age1    0.018022   0.008031   2.244   0.0263
## IllnessDurBinhealthy:age1  0.007988   0.007181   1.112   0.2678
## 
## Residual standard error: 0.4976 on 148 degrees of freedom
## Multiple R-squared:  0.4278, Adjusted R-squared:  0.4085 
## F-statistic: 22.13 on 5 and 148 DF,  p-value: < 2.2e-16

plot(allEffects(m2), grid = T)
```



9) (One last time!): 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

10) 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.