Lab 4

9/11/2024

Lab 4 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 *selecting the group members in grade scope* 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)),
             IllnessDurBin = fct_recode(IllnessDurBin,
                                        Low = "[1,16]",
                                        High = "(16,60]",
                                        Healthy = "(Missing)"),
             logNormVolume = log(NormVolume)
             )

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)
```

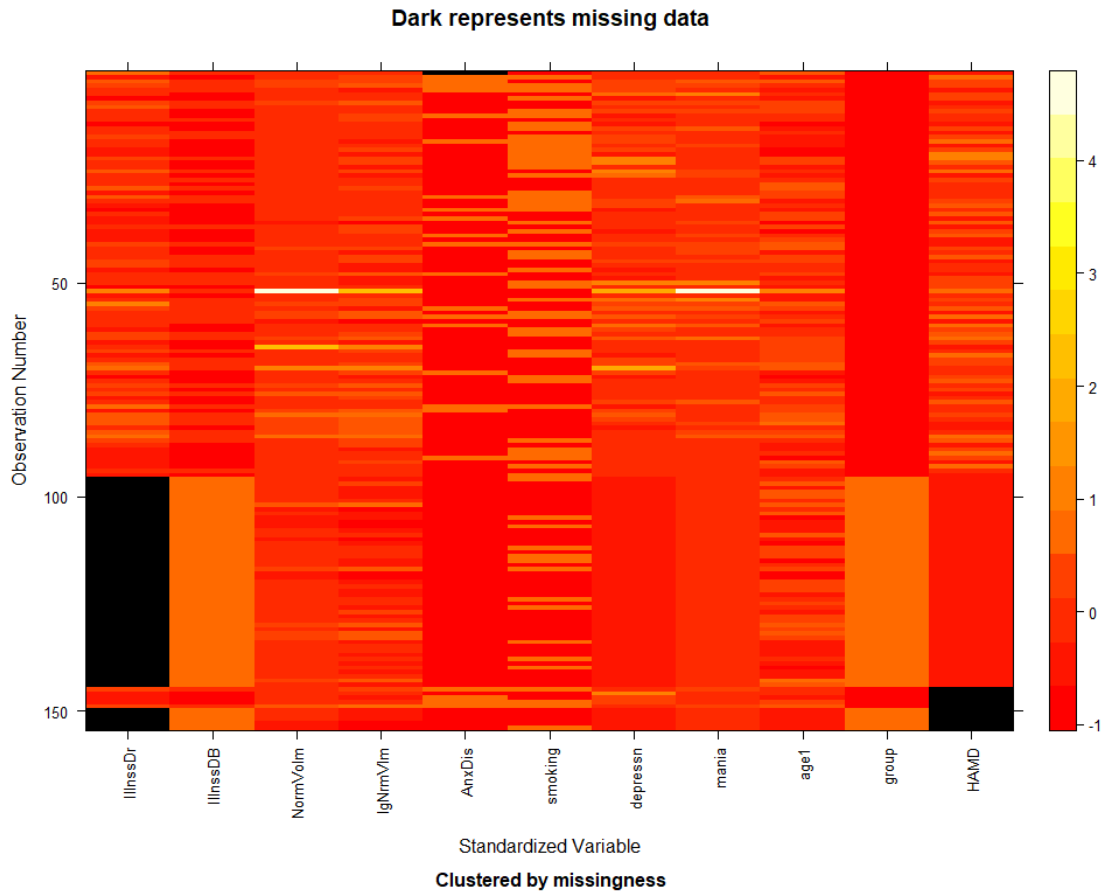
```
##   IllnessDurBin min Q1 median    Q3 max      mean        sd  n missing
## 1           Low   1  5      9 13.00  16  9.096154  4.827453 52       0
## 2          High  17 20     29 35.25  60 29.854167 10.595040 48       0
## 3       Healthy  NA NA     NA    NA  NA       NaN        NA  0      54
```

In lab 3, we created `IllnessDurBin` that viewed missing illness duration as a level of a factor that relates to healthy subjects but we didn't fully explore missingness in the data set. One tool for doing that is using the `mi` package's `missing_data.frame`. It requires the object passed to it to be `data.frame`, so it is good practice to use `as.data.frame` on data sets before using it as the function will provide a cryptic error message if you pass a `tibble` to the function. The following code picks some of the variables we will be using below and then runs the function on the data.frame to create what I called `mbpdata_select`. Then it generates a tally of different missingness patterns and a plot of the results. In the image (tile-plot), black bars are missing and the color-ramp tries to explore values on the predictors (even if they are categorical), with the observations sorted based on their missingness patterns.

```
library(mi)
bpdata_select <- bpdata %>% dplyr::select(IllnessDuration, IllnessDurBin,
                                  NormVolume, logNormVolume, AnxDis,
                                  smoking, depression,
                                  mania, age1,
                                  group, HAMD)

mbpdata_select <- bpdata_select %>% as.data.frame() %>% missing_data.frame()
table(mbpdata_select@patterns)
```

```
##
##               nothing                  HAMD                AnxDis
##                    94                     5                     1
##       IllnessDuration IllnessDuration, HAMD
##                    49                     5
```

```
image(mbpdata_select)
```



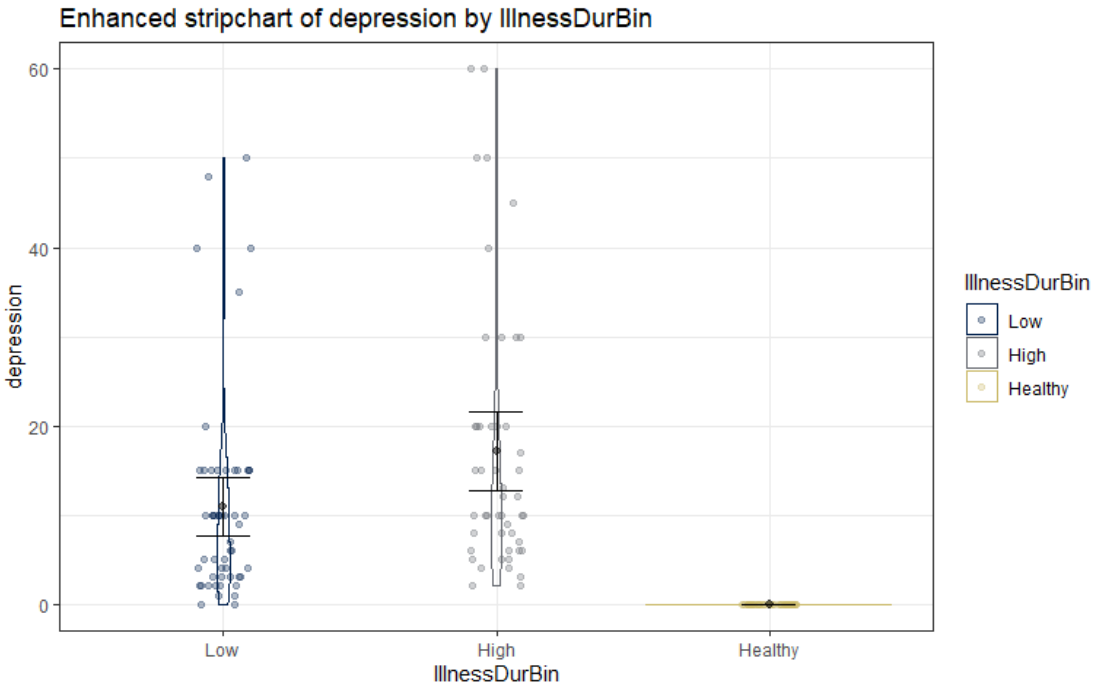
1) How many missing data patterns does the `table` result suggest for these observations? **There five identified patterns of missing data.**

2) Explain the patterns and coloring in the two illness duration variables. **This is a categorical column and the plot standardized this as numeric. The orange of IllnssDB matches that of the black, or missing, pattern in IllnssDr. This makes sense since in the uncleaned version, healthy patients were represented with an NA.**

3) We will work with a smaller set of variables to explore the `depression` response, which is a count of depressive events, and age, illness duration, and group. We are not told the time scale on this count, but it looks like it could be a lifetime count. In these variables, there are no missing values, so we can proceed with using all the observations for the following data visualization and models. In the enhanced stripchart and `favstats` output that is provided, what is the issue with modeling the count of depression events based on `IllnessDurBin`?

```
bpdata2 <- bpdata %>% dplyr::select(age1, IllnessDurBin, group, depression)

mbpdata2 <- bpdata2 %>% as.data.frame() %>% missing_data.frame()
table(mbpdata2@patterns)
```

```
##
## nothing
##     154

favstats(depression ~ IllnessDurBin, data = bpdata2)

##   IllnessDurBin min   Q1 median Q3 max     mean       sd  n missing
## 1           Low   0 3.00    9.5 15  50 10.98077 11.73313 52       0
## 2          High   2 6.75   11.0 20  60 17.14583 15.05308 48       0
## 3       Healthy   0 0.00    0.0  0   0  0.00000  0.00000 54       0

enhanced_stripchart(depression ~ IllnessDurBin, data = bpdata2)
```



All healthy patients were never depressed, and thus the relationship is negligible. If we were to model this relationship, the interpretation would be suspect, if not erroneous.

4) The following code removes the wrong group from the data set. Fix it to remove the healthy subjects from the data set.

```
bpdata3 <- bpdata2 %>% dplyr::filter(IllnessDurBin != "Healthy") %>%
  mutate(IllnessDurBin = factor(IllnessDurBin)) #Cleans out unused level of
the factor from before
favstats(depression ~ IllnessDurBin, data = bpdata3)

##   IllnessDurBin min   Q1 median Q3 max     mean       sd  n missing
## 1           Low   0 3.00    9.5 15  50 10.98077 11.73313 52       0
## 2          High   2 6.75   11.0 20  60 17.14583 15.05308 48       0
```

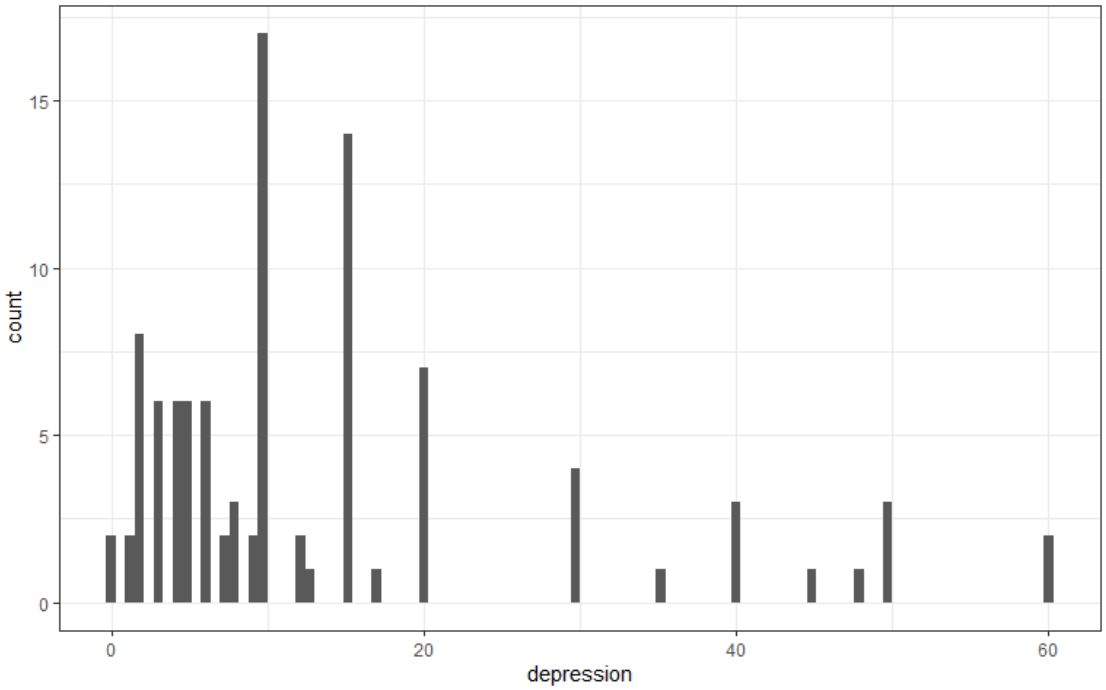
5) Review the values in `depression` in your new data set `bpdata3` by exploring the tally of number of observations at each value and the histogram of the observations. Does

this meet the criteria to be possibly modeled using a Poisson distribution? Why is this not a grouped binomial response?

```
tally(depression ~ 1, data = bpdata3)
```

```
##            1
## depression  1
##          0  2
##          1  2
##          2  8
##          3  6
##          4  6
##          5  6
##          6  6
##          7  2
##          8  3
##          9  2
##          10 17
##          12  2
##          13  1
##          15 14
##          17  1
##          20  7
##          30  4
##          35  1
##          40  3
##          45  1
##          48  1
##          50  3
##          60  2
```

```
bpdata3 %>% ggplot(aes(x = depression)) +
  geom_histogram(bins = 100)
```



These data follow a poisson distribution because they are continuous count data. The zeros in the poisson distribution are OK. It is not a grouped binomial distribution because there is no m group of bins/ observations.

6) Continue to use `bpdata3` to fit an appropriate `glm` for the depression response with `age1` and `IllnessDurBin` as predictors (no interaction). Generate a model `summary` for the model and run `confint` on the model. No discussion.

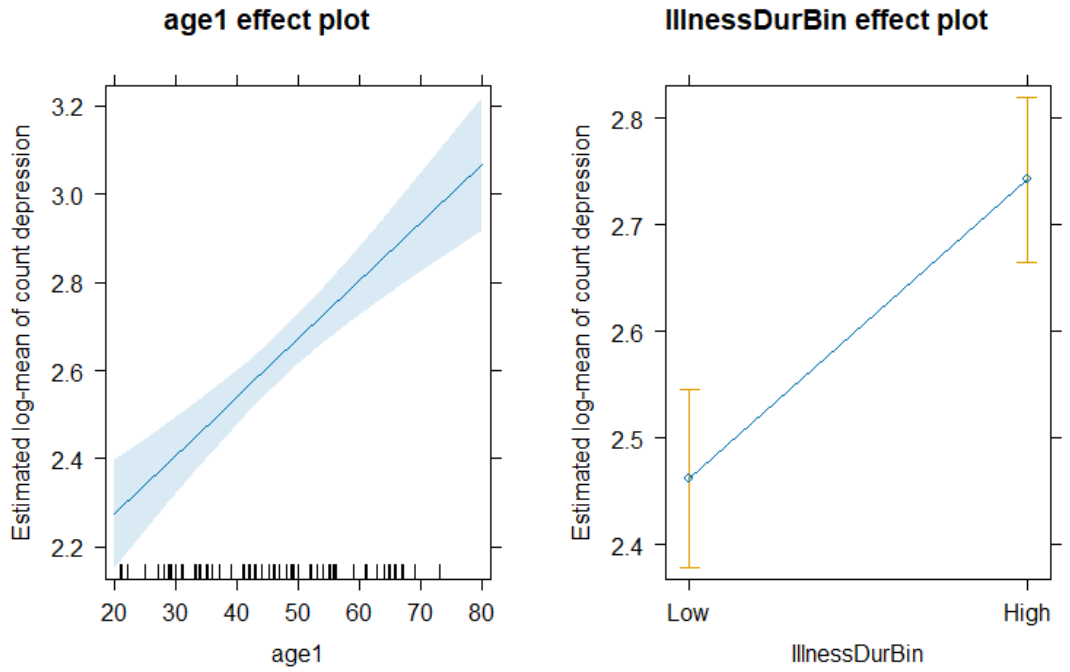
```
m1 <- glm(depression ~ age1 + IllnessDurBin, data = bpdata3, family =
poisson)
summary(m1)

##
## Call:
## glm(formula = depression ~ age1 + IllnessDurBin, family = poisson,
##     data = bpdata3)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.877118   0.095227  19.712  < 2e-16
## age1              0.013213   0.002123   6.223 4.88e-10
## IllnessDurBinHigh 0.280017   0.060679   4.615 3.94e-06
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1103.86  on 99  degrees of freedom
## Residual deviance:  997.21  on 97  degrees of freedom
## AIC: 1405.6
##
## Number of Fisher Scoring iterations: 5
```

```
confint(m1)
```

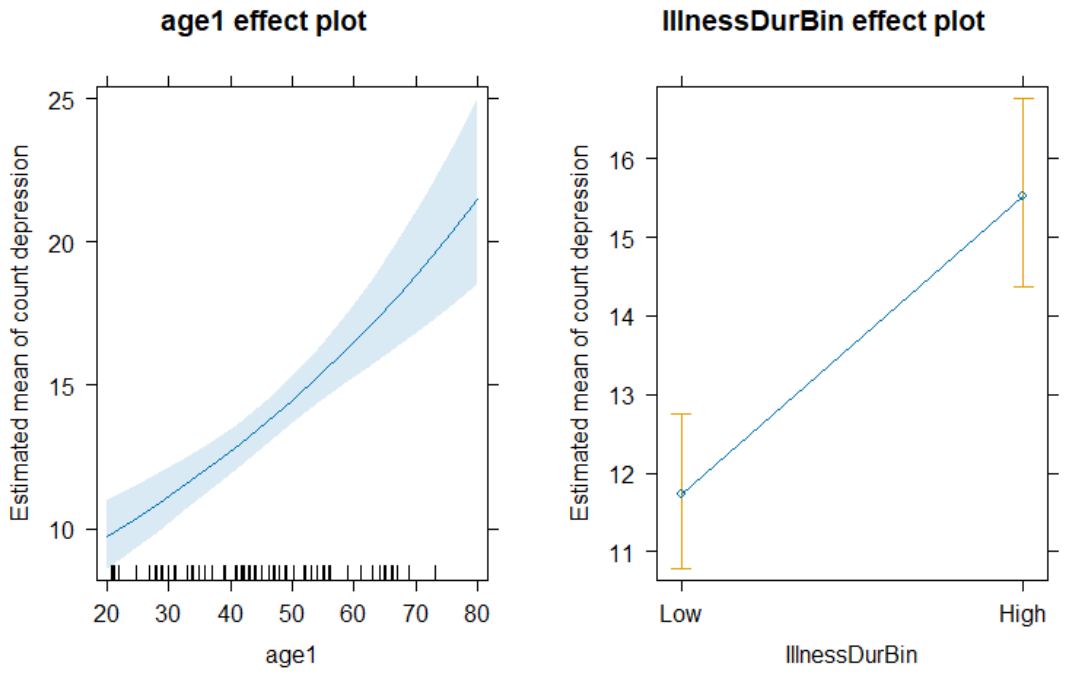
```
##                          2.5 %     97.5 %
## (Intercept)       1.689292062 2.06260076
## age1              0.009047565 0.01737163
## IllnessDurBinHigh 0.161338859 0.39923651
```

7) Generate effects plots on both the link and response scales with the grid added to each. Improve the y-axis labels similar to what I did in the lecture notes. Note: if your plots don't change between link and response, then you likely have an issue in the model you fit.

```
plot(allEffects(m1),type = 'link',  ylab = "Estimated log-mean of count depression")
```




```
plot(allEffects(m1), type ='response', ylab = "Estimated mean of count depression")
```



8) Write up a "size" interpretation for `age1` in the previous model in the link (log-mean) scale.

For two otherwise similar patients, for every 1 year increase in age, we estimated a log-mean increase in depression of 0.013 (95%:0.009, 0.017) than that of the other patient, controlling for Illness Duration Bin.

9) Write up a size interpretation for `age1` in the previous model on the response (mean) scale.

For two otherwise similar patients, for every 1 year increase in age, we estimated that the mean depression count increased by exp(0.013) = 1.01 (95%:exp(0.009), exp(0.017)) than that of the other patient, controlling for Illness Duration Bin.

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.

NONE