HW 2

Ceili DeMarais

Due 1/31/2025

HW 2 Instructions:

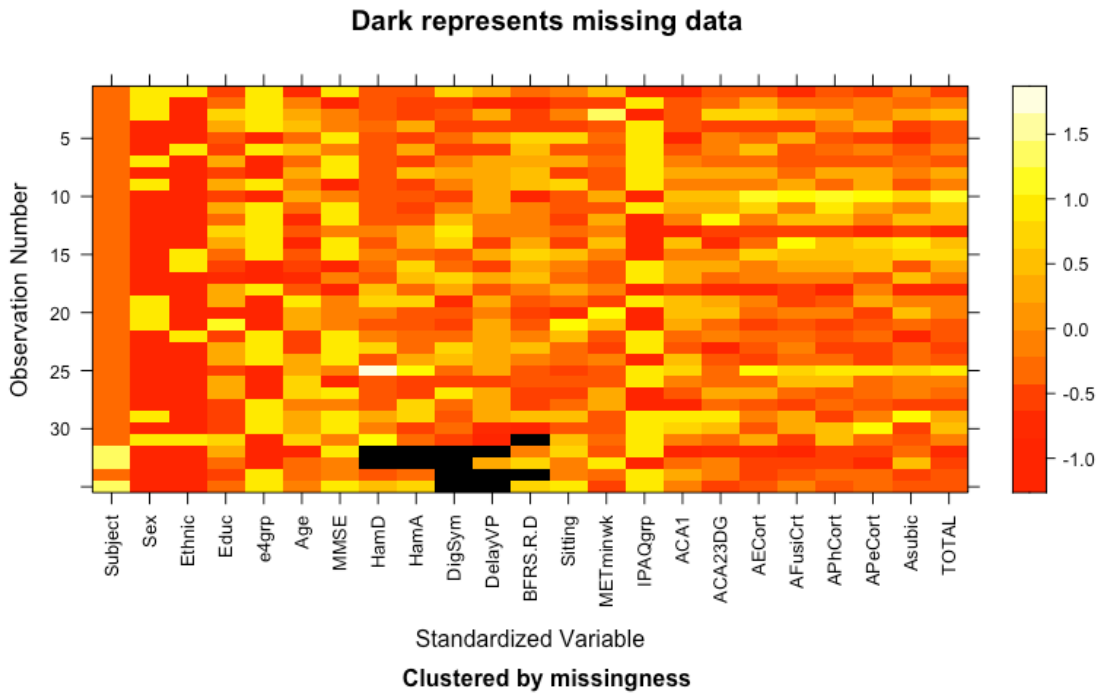
Continuing with sitting and brain impacts and working individually, complete the following based on the Siddarth et al. (2018) paper.

- Siddarth P, Burggren AC, Eyre HA, Small GW, Merrill DA (2018) Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults. PLoS ONE 13(4): e0195549. https://doi.org/10.1371/journal.pone.0195549

1) The following code checks for missing values in the data set. Generate a "clean" data set with no missing observations using `drop_na()` but also keep a version with all observations. What is the sample size to start with and after "cleaning"?

- **The beginning sample size was 35 observations, and after cleaning this number dropped to 30 observations.**

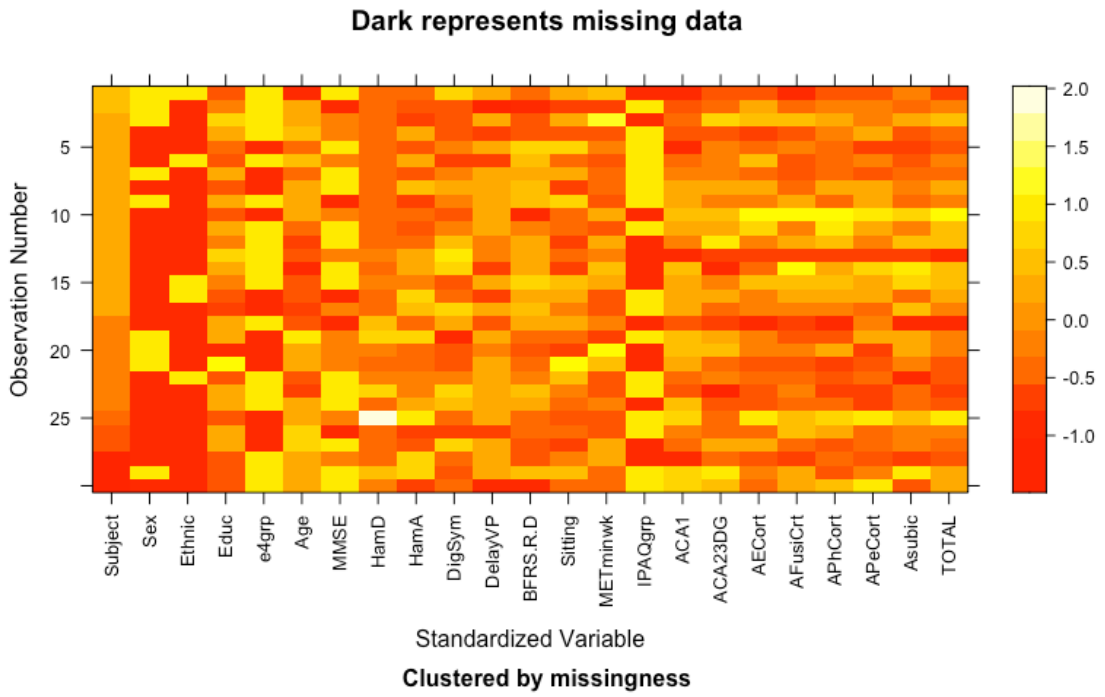
```
data(sit_and_brain)
library(mi)
tdf <- missing_data.frame(data.frame(sit_and_brain))

## 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,] "HamD" "HamA"

image(tdf)
```




```
table(tdf@patterns)
```

```
##
##                                          nothing
##                                               30
##                   BFRSelective.Reminding.Delayed
##                                                1
##                                  DigSym, DelayVP
##                                                1
## DigSym, DelayVP, BFRSelective.Reminding.Delayed
##                                                1
##                               HamD, HamA, DigSym
##                                                1
##                     HamD, HamA, DigSym, DelayVP
##                                                1
```

```
dat <- sit_and_brain %>% drop_na()
tdf1 <- missing_data.frame(data.frame(dat))
image(tdf1)
```



2) For this question, code is provided to make a version of Table 1 that includes all the variables in the data set except for `Subject` using the `datasummary_balance` function from the `modelsummary` package. Your task is to create a second version of Table 1 for the the "cleaned" version of the data set from the previous question. No discussion.

- Note that in contrast to the version in my notes, use `~1` as the formula in order to not split the results up by groups, so it will mostly match their version of the table. You do not need to work on variable labels/names or sorting to match their version of the results.

```
library(modelsummary)

datasummary_balance(~1, data = sit_and_brain %>% dplyr::select(-Subject),
                    stars = T,
                    dinm = T)
```

	Mean	Std. Dev.
Educ	16.4	2.5
Age	60.4	8.1
MMSE	29.3	0.7
HamD	1.8	3.0
HamA	4.4	3.6
DigSym	66.9	16.2
DelayVP	6.6	2.2

		Mean	Std. Dev.
BFRSelective Reminding Delayed		7.9	3.3
Sitting		7.2	3.3
METminwk		1521.3	1225.7
ACA1		2.0	0.1
ACA23DG		2.9	0.2
AECort		2.6	0.3
AFusiCort		2.7	0.2
APhCort		2.8	0.4
APeCort		2.6	0.3
Asubic		2.1	0.2
TOTAL		2.5	0.2
		N	Pct.
Sex	F	25	71.4
	M	10	28.6
Ethnic	Caucasian	29	82.9
	Other	6	17.1
e4grp	E4	15	42.9
	Non-E4	20	57.1
IPAQgrp	High	14	40.0
	Low	21	60.0

```
datasummary_balance(~1, data = dat %>% dplyr::select(-Subject),
                    stars = T,
                    dinm = T)
```

		Mean	Std. Dev.
Educ		16.3	2.5
Age		60.6	8.1
MMSE		29.3	0.7
HamD		1.5	2.8
HamA		4.3	3.6
DigSym		67.6	16.0
DelayVP		6.7	2.0
BFRSelective Reminding Delayed		7.6	3.3
Sitting		6.7	3.2
METminwk		1530.0	1218.4

		Mean	Std. Dev.
ACA1		2.0	0.1
ACA23DG		3.0	0.2
AECort		2.6	0.3
AFusiCort		2.7	0.2
APhCort		2.8	0.4
APeCort		2.6	0.3
Asubic		2.1	0.2
TOTAL		2.6	0.2
		N	Pct.
Sex	F	21	70.0
	M	9	30.0
Ethnic	Caucasian	25	83.3
	Other	5	16.7
e4grp	E4	12	40.0
	Non-E4	18	60.0
IPAQgrp	High	13	43.3
	Low	17	56.7

3) Compare the two sets of summary statistics (i.e., your two tables) to their Table 1 for Age. Does it appear that they used all available information or only "complete cases" (this is what you get if you run `drop_na()` on the entire data set to get only observations that are "complete" on all considered variables)?

- **It looks as if they used all available information in their Table 1. Their sample size was 35 and the mean age, 60.4, was the same as our "non-cleaned" dataset with NA values.**

4) In their Table 2, how many p-values do they report? What was their rule for "bolding" p-values in that table?

- **They report 16 p-values, accounting for each of their beta coefficients. They bolded p-values that were less than or equal to 0.05. These values they considered "s-word" in their write up.**

5) (Re)-read section 1.8 in Greenwood (202X) that is titled "Reproducibility Crisis: Moving beyond $p < 0.05$, publication bias, and multiple testing issues". Using a result from that section, what is the probability of at least one Type I error in their

Table 2 if the tests were all independent of each other and all null hypotheses were true and they used a 0.05 cutoff for each test?

- **The probability of there being at least one Type I error in their Table 2 if all the tests were independent of each other and all null hypotheses were true and they used a 0.05 cutoff for each test is 0.5599 (55.99%).**

```
1-(0.95^16)
```

```
## [1] 0.5598733
```

6) The Bonferroni correction involves taking the p-values and multiplying the p-values by the number of tests being considered (adjusted p-values cannot exceed 1 since this is a probability, so it is really $p_{adj} = min(p - value * numberoftests, 1)$). The vector of bolded p-values is provided below. Use that vector to generate the Bonferroni-adjusted p-values for those four tests (this should account for the total number of tests in the table, not just those four). What does this do to your assessment of the potential for evidence against the suite of null hypotheses that sitting and activity are not related to the various brain thickness measurements, controlled for the other focal variable (sitting or activity) and the age of the subjects?

- **This decreases the evidence against the suite of null hypotheses that sitting and activity are not related to the various brain thickness measurements, controlled for the other focal variable (sitting or activity) and the age of the subjects. Before the adjustment there was strong evidence against the null, but after adjustment there is little to none**.

```
pvals <- c(0.03, 0.05, 0.007, 0.04)

pvals_adj <- p.adjust(pvals, method = "bonferroni", n = 16)

data.frame(Original_pval = pvals, Adjusted_pval = pvals_adj)
```

```
##   Original_pval Adjusted_pval
## 1         0.030         0.480
## 2         0.050         0.800
## 3         0.007         0.112
## 4         0.040         0.640
```

Use the full data set for this (`sit_and_brain`) and the remaining questions:

7) They note that they log-transformed the Physical activity (`METminwk`) variable but don't state which type of log transformation they used. Create three versions of the log of the METminwk: natural `log`, `log10`, and `log2` using `mutate`. Fit three models for `TOTAL` with each including `Sitting` and Age and then try each of the versions of the log-METminwk. Generate model summaries and `confint` results. Which of three

sets of results seem to match their reported results for total MTL thickness best that are reported in the first two rows of Table 2?

- Note that again they might not have used proper rounding rules to report their results.
    - **Their reported results for total MTL thickness are -0.02 (-0.04, -0.002) for sitting, and 0.007 (-0.07, 0.08) for physical activity. Based on the comparison of different results below, it looks like they used the natural log.**

```
sit_and_brain <- sit_and_brain %>% mutate(METminwk_log = log(METminwk),
                                          METminwk_log10 = log10(METminwk),
                                          METminwk_log2 = log2(METminwk))

m_log <- lm(TOTAL ~ Sitting + Age + METminwk_log, data = sit_and_brain)
summary(m_log)

##
## Call:
## lm(formula = TOTAL ~ Sitting + Age + METminwk_log, data = sit_and_brain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30673 -0.13091 -0.02548  0.13243  0.44786
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.363491   0.402649   5.870 1.78e-06
## Sitting      -0.020952   0.009500  -2.206    0.035
## Age           0.004535   0.004022   1.127    0.268
## METminwk_log  0.006911   0.035698   0.194    0.848
##
## Residual standard error: 0.1811 on 31 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1117
## F-statistic: 2.425 on 3 and 31 DF,  p-value: 0.08441

confint(m_log)

##                    2.5 %       97.5 %
## (Intercept)   1.542283379  3.184698426
## Sitting      -0.040326019 -0.001577243
## Age          -0.003668899  0.012738963
## METminwk_log -0.065895046  0.079717746

m_log10 <- lm(TOTAL ~ Sitting + Age + METminwk_log10, data = sit_and_brain)
summary(m_log10)

##
## Call:
## lm(formula = TOTAL ~ Sitting + Age + METminwk_log10, data = sit_and_brain)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30673 -0.13091 -0.02548  0.13243  0.44786
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.363491   0.402649   5.870 1.78e-06
## Sitting        -0.020952   0.009500  -2.206    0.035
## Age             0.004535   0.004022   1.127    0.268
## METminwk_log10  0.015914   0.082198   0.194    0.848
##
## Residual standard error: 0.1811 on 31 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1117
## F-statistic: 2.425 on 3 and 31 DF,  p-value: 0.08441
```

```
confint(m_log10)
```

```
##                       2.5 %       97.5 %
## (Intercept)     1.542283379  3.184698426
## Sitting        -0.040326019 -0.001577243
## Age            -0.003668899  0.012738963
## METminwk_log10 -0.151728951  0.183556893
```

```
m_log2 <- lm(TOTAL ~ Sitting + Age + METminwk_log2, data = sit_and_brain)
summary(m_log2)
```

```
##
## Call:
## lm(formula = TOTAL ~ Sitting + Age + METminwk_log2, data = sit_and_brain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30673 -0.13091 -0.02548  0.13243  0.44786
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.363491   0.402649   5.870 1.78e-06
## Sitting       -0.020952   0.009500  -2.206    0.035
## Age            0.004535   0.004022   1.127    0.268
## METminwk_log2  0.004791   0.024744   0.194    0.848
##
## Residual standard error: 0.1811 on 31 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1117
## F-statistic: 2.425 on 3 and 31 DF,  p-value: 0.08441
```

```
confint(m_log2)
```

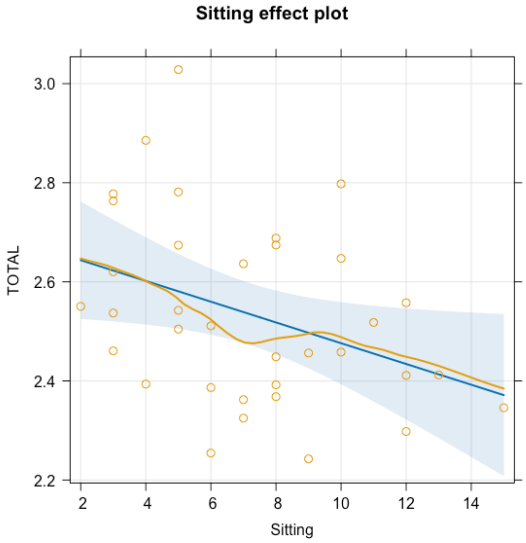
```
##                       2.5 %       97.5 %
## (Intercept)    1.542283379  3.184698426
## Sitting       -0.040326019 -0.001577243
```

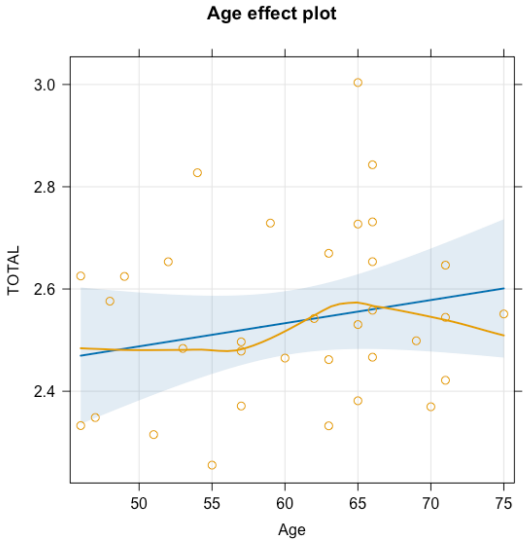
```
## Age           -0.003668899  0.012738963
## METminwk_log2 -0.045674965  0.055256131
```

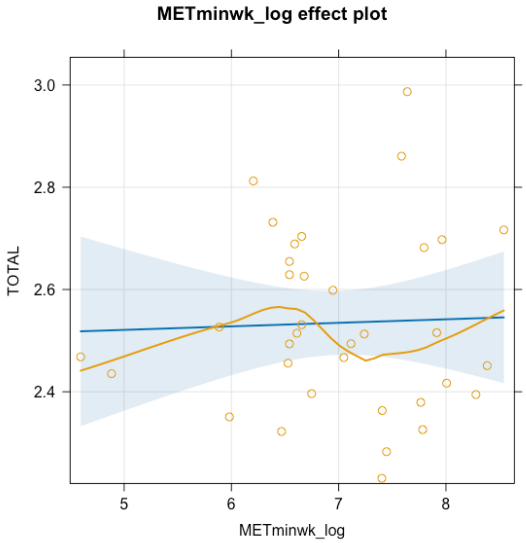
8) Make an effects plot for the model that best matches their results from the previous question. Based on the effects plots, which of the three predictors had the most impact on the estimated mean response? Use the `grid=T` option to help with assessing change over the range in the estimated mean response across the range of observed predictor values.

- **Based on the effects plots below, it seems as though the sitting predictor had the most impact on the estimated mean response.**

```
plot(allEffects(m_log, residuals = T), grid = T)
```







9) I want to check your version of R. Do not edit the code below as it will check your version of R and report that in the knitted document. You should be using R 4.4.2 and you should make sure your packages are relatively up to date.

- R version (short form): 4.4.2

10) Document any collaborations or pertinent discussions with other students or resources outside of your group and the resources that I am providing and the *Sleuth* that you used to complete this assignment *or report that you did not have any*. If you used generative AI (chatGPT, Bard, etc.), report which question(s) and how/what you asked for.

NONE