HW 1

Ceili DeMarais

Due Wednesday, January 22 at 9 am

# HW 1 Instructions:

You may discuss the work with other students but should complete this assignment in an independent fashion (you should try the code and writing on your own but can benefit from discussions with other students). At the end of the assignment you will document any discussions that have a substantive impact on your results. Also document any resources you used outside of those that I provide, including things like ChatGPT.

Answer each question inline after the prompts with pertinent code, results, and graphs preceding your written answer.

Make sure to run spell-check using the “ABC check” button near the filename or use the text underlining suggestions in the document. Knit this .Rmd to word, then convert the word document to a PDF. Submit just the PDF for grading in gradescope.com HW 1.

**Uncomment the lines of code above to install ggResidpanel and catstats2 from my github repository. Then re-comment them for all following use of the file.** Contact us ASAP if you have difficulties installing and loading the packages or finding the data set. The most critical package for this assignment is catstats2.

After knitting this document to a Word document, save as or print to PDF to create a PDF of the file. Submit this document to gradescope.com and use the interface to identify the pages that contain the questions and your answers for each numbered question.

## MS Teams Access:

1. MS Teams OneNote access: It is critical that you gain access to OneNote from MS teams that is tied to our MS Teams channel to access lecture notes, daily announcements (provided near the starting point for each lecture day), and feedback on projects. To prove you have access and help us to learn who you are, please post a selfie/picture of you where you are identifiable in your Homework - HW 1 Selfie page in OneNote. Only the instructors will see the picture. No discussion, just the work in OneNote.

## Sitting and brain size… initial exploration:

* Read 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>
* The data were downloaded from <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0195549> and are provided in catstats2 with data(sit\_and\_brain) with some information about the data set via help(sit\_and\_brain).
* Note that the questions are inspired by the practice problems in <https://bookdown.org/roback/bookdown-BeyondMLR/ch-MLRreview.html#introduction-to-beyond-multiple-linear-regression>

Siddarth et al. (2018) researched the relationship between time spent sitting (sedentary behavior) and the thickness of a participant’s medial temporal lobe (MTL) in a 2018 paper entitled, “Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults”. In other research, MTL volume has been found to be negatively associated with Alzheimer’s disease and memory impairment. Key variables measured in this study include:

* TOTAL = Medial temporal lobe thickness in mm
* Sitting = Reported hours/day spent sitting
* METminwk = Reported metabolic equivalent unit minutes per week
* Age = Age in years
* Sex = Sex (M = Male, F = Female)
* Educ = Years of education completed

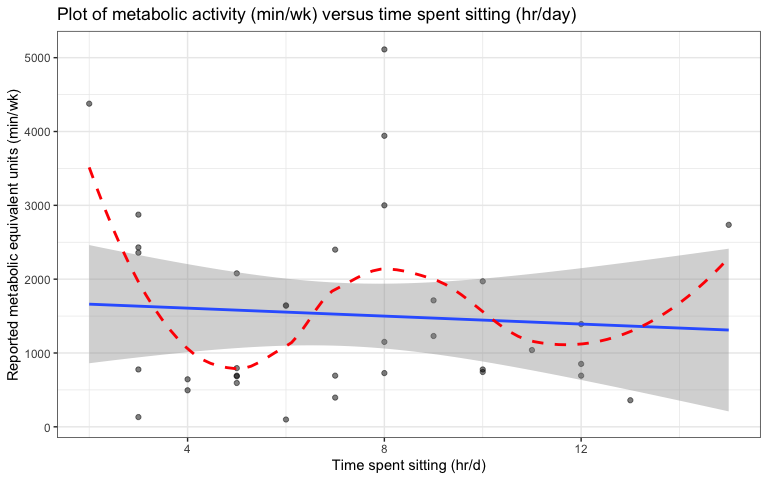
library(catstats2)  
data(sit\_and\_brain)  
glimpse(sit\_and\_brain)

## Rows: 35  
## Columns: 23  
## $ Subject <dbl> 9690, 9722, 9735, 9787, 10010, 10021,…  
## $ Sex <chr> "M", "M", "F", "F", "F", "F", "F", "F…  
## $ Ethnic <chr> "Caucasian", "Other", "Caucasian", "C…  
## $ Educ <dbl> 14, 20, 14, 14, 18, 18, 14, 18, 18, 1…  
## $ e4grp <chr> "Non-E4", "E4", "Non-E4", "Non-E4", "…  
## $ Age <dbl> 66, 71, 66, 63, 71, 71, 66, 65, 51, 5…  
## $ MMSE <dbl> 30, 29, 30, 29, 30, 28, 29, 30, 30, 3…  
## $ HamD <dbl> 4, 8, 2, 0, 0, 0, 13, 0, 6, 2, 0, 2, …  
## $ HamA <dbl> 9, 4, 0, 9, 1, 0, 12, 7, 4, 4, 2, 2, …  
## $ DigSym <dbl> 57, 47, 60, 64, 94, 48, 62, 81, 92, 6…  
## $ DelayVP <dbl> 8, 1, 3, 8, 8, 4, 8, 8, 8, 8, 8, 7, 8…  
## $ `BFRSelective Reminding Delayed` <dbl> 11, NA, 2, 4, 5, 6, 6, 9, 11, 8, 6, 4…  
## $ Sitting <dbl> 10, 11, 5, 7, 3, 5, 4, 6, 7, 10, 15, …  
## $ METminwk <dbl> 777.0, 1039.8, 795.0, 2400.0, 2358.0,…  
## $ IPAQgrp <chr> "Low", "Low", "Low", "High", "High", …  
## $ ACA1 <dbl> 2.25275, 2.08825, 2.20960, 1.82060, 1…  
## $ ACA23DG <dbl> 3.36390, 2.91600, 3.15045, 2.86030, 3…  
## $ AECort <dbl> 2.63580, 2.77730, 2.40835, 2.30295, 2…  
## $ AFusiCort <dbl> 2.78055, 2.49820, 2.76160, 2.48635, 2…  
## $ APhCort <dbl> 2.46110, 3.13505, 3.17635, 2.69160, 2…  
## $ APeCort <dbl> 2.61820, 2.56485, 3.28765, 2.37405, 2…  
## $ Asubic <dbl> 2.58200, 1.99050, 1.88995, 1.85835, 2…  
## $ TOTAL <dbl> 2.67061, 2.56716, 2.69771, 2.34203, 2…

1. In their article’s introduction, Siddarth et al. differentiate their analysis on sedentary behavior from an analysis on active behavior by citing evidence supporting the claim that, “one can be highly active yet still be sedentary for most of the day”, suggesting that the two potential predictor variables might not be related to each other. Make a scatterplot of the METminwk and Sitting using ggplot and add linear and nonparametric smoothing lines. No discussion.

* Make sure to add a title to the figure and modify the axis labels to include the units of the variables (in parentheses) and make the plot so you can see any points that are overplotting (either using alpha or jittering or both).

sit\_and\_brain %>% ggplot(mapping = aes(y = METminwk, x = Sitting)) +   
 geom\_point(alpha = 0.5) +   
 geom\_smooth(method = "lm") +   
 geom\_smooth(col = "red", lty = 2, se = F) +  
 labs(x = "Time spent sitting (hr/d)",  
 y = "Reported metabolic equivalent units (min/wk)",   
 title = "Plot of metabolic activity (min/wk) versus time spent sitting (hr/day)")



1. Fit a linear model with METminwk and Sitting as your explanatory and response variables where the variable on the y-axis in your previous plot is the response variable for this model. Calculate and report the R-squared for this model in a sentence. Then use the R-squared result (no tests or other measures) to discuss the claim that sedentary behaviors are not related to physical activity *in these data*.

* The R-squared value 0.005345 suggests that 0.5345% of the variation in METminwk is explained by the model that contains time sitting. This very small percentage suggests very weak evidence to support the claim of a relation between sedentary behavior and physical activity.

model <- lm(METminwk ~ Sitting, data = sit\_and\_brain)  
summary(model)

##   
## Call:  
## lm(formula = METminwk ~ Sitting, data = sit\_and\_brain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1502.5 -872.5 -379.0 624.4 3612.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1715.37 506.43 3.387 0.00184  
## Sitting -26.96 64.02 -0.421 0.67641  
##   
## Residual standard error: 1241 on 33 degrees of freedom  
## Multiple R-squared: 0.005345, Adjusted R-squared: -0.0248   
## F-statistic: 0.1773 on 1 and 33 DF, p-value: 0.6764

1. Fit a preliminary SLR model with TOTAL as the response and Sitting as the sole explanatory variable. Summarize the “size” using the predictive version of the slope interpretation and “evidence” (report the t-test result) for the predictor in this model. See Greenwood (202X) Section 1.9.1 and the review lecture notes and related recording for more on reporting these two aspects of the slope in this model.

slr1 <- lm(TOTAL ~ Sitting, data = sit\_and\_brain)  
summary(slr1)

##   
## Call:  
## lm(formula = TOTAL ~ Sitting, data = sit\_and\_brain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33511 -0.13432 -0.00252 0.11527 0.46907   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.69951 0.07309 36.933 <2e-16  
## Sitting -0.02288 0.00924 -2.476 0.0186  
##   
## Residual standard error: 0.1791 on 33 degrees of freedom  
## Multiple R-squared: 0.1567, Adjusted R-squared: 0.1312   
## F-statistic: 6.132 on 1 and 33 DF, p-value: 0.01857

* **Evidence answer: There is strong evidence against the null hypothesis of no linear relationship between Sitting and Total brain size (*t*33​=−2.5, two-sided p-value = 0.0186), so we would conclude that there is a relationship between them and keep the sitting time as a predictor for total brain size.**
* **Size answer: For two subjects that are one hour of sitting different but otherwise similar, the estimated mean brain size is -0.023 mm different for the longer sitting one than the shorter sitting one (95% confidence interval: -0.042 to -0.004).**
* **Or could switch sign use words for the direction (and then also switch the sign and order on the CI): For two subjects that are one hour of sitting different but otherwise similar, the estimated mean brain size is 0.023 mm lower for the longer sitting one than the shorter sitting one (95% confidence interval: 0.004 to 0.042).**

1. Now fit a linear model for TOTAL that contains Sitting and METminwk (no interaction). Use the estimated model to predict the TOTAL for a subject that sits for 14 hours per day and has a MET of 5000 minutes per week (show your code and work for doing this).

mlr1 <- lm(TOTAL ~ Sitting + METminwk, data = sit\_and\_brain)  
summary(mlr1)

##   
## Call:  
## lm(formula = TOTAL ~ Sitting + METminwk, data = sit\_and\_brain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33495 -0.13171 -0.00527 0.11380 0.47001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.703e+00 8.616e-02 31.369 <2e-16  
## Sitting -2.293e-02 9.408e-03 -2.438 0.0205  
## METminwk -1.886e-06 2.551e-05 -0.074 0.9415  
##   
## Residual standard error: 0.1818 on 32 degrees of freedom  
## Multiple R-squared: 0.1569, Adjusted R-squared: 0.1042   
## F-statistic: 2.976 on 2 and 32 DF, p-value: 0.06523

2.703 - (0.02293\*14) - (0.000001886\*5000)

## [1] 2.37255

* We predict that a subject that sits for 14 hours a day and has a MET of 5000 minutes per week, will have a Medial temporal lobe thickness of 2.37255 mm.

1. Select **all** of the following that apply related to the previous model and prediction (Note in answer which letters you are selecting - no discussion).

**a) The prediction is problematic because the combination of the two predictors is not actually possible in a week based on a week having 24 hours 60 minutes 7 days = 10,080 minutes.**

**b) The prediction is problematic because it is extrapolation since the combination of the two predictors is far from the combinations of the two predictors observed in the data set, as seen in the scatterplot.**

1. The prediction is not problematic because it is not extrapolation since the value of each of the predictors was observed in the data set, as seen in the scatterplot.
2. Because we used least squares estimation to find the slope coefficients in the model, it is reasonable to generate predictions from the model since it minimized the squared error to find the slope coefficients.
3. Based on the two-sided p-value for METminwk from the t-test in the model, we would reject the null hypothesis of METminwk being related to the response and would keep it in the model.
4. Based on the last row of the model summary output with F(2,32) = 2.976 and a p-value of 0.065, we should remove the METminwk from the model because the p-value is not significant.

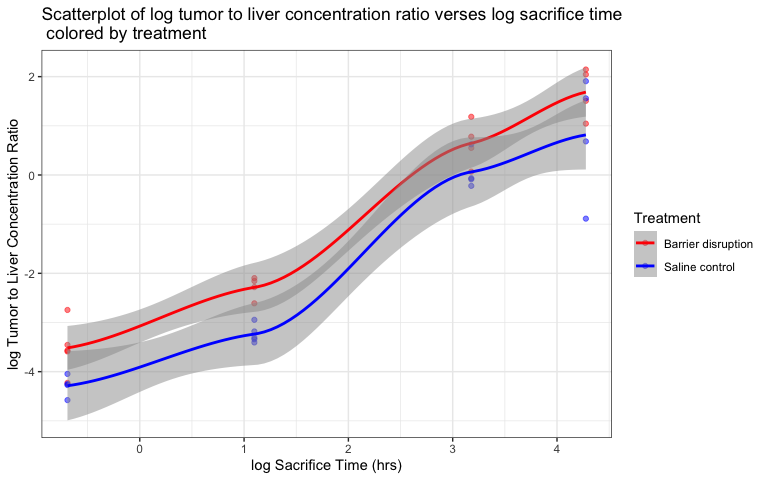
**Chosen letters are bolded.**

1. Write a *Scope of Inference* for this study for the model that contains Sitting and METminwk for TOTAL. See the related discussion in Greenwood (202X) Section 1.7.

* We cannot generalize inference to the population of interest because the sample was not random and could be biased. We cannot establish causal inference due to lack of random assignment of the treatment.

1. In the Statistical Sleuth Chapter 11, Display 11.5 shows a plot of two variables from the case1102 data set from the Sleuth3 R package. The following improves on the code in the help file for case1102 using mutate to make the variables needed for the plot. Use ggplot to make a plot similar to Display 11.5 but that either jitters or uses transparency (alpha) for the points displayed and adds nonparametric smoothing lines for each Treatment group. You do not need the “fancy” double axis tick marks they used in the book, this can just be on the log-scale.

library(dplyr)  
data(case1102)  
case1102 <- case1102 %>% mutate(ratio = Brain/Liver,  
 logTime = log(Time),  
 logratio = log(ratio))  
  
case1102 %>% ggplot(mapping = aes(x = logTime, y = logratio)) +  
 geom\_point(mapping = aes(col = Treatment), alpha = 0.5) +  
 geom\_smooth(mapping = aes(col = Treatment), method = "loess", alpha = 0.5) +  
 scale\_color\_manual(values = c("BD" = "red", "NS" = "blue"),  
 labels = c("Barrier disruption", "Saline control")) +  
 labs(x = "log Sacrifice Time (hrs)",  
 y= "log Tumor to Liver Concentration Ratio",  
 title = "Scatterplot of log tumor to liver concentration ratio verses log sacrifice time \n colored by treatment")



1. 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

1. Document any collaborations or pertinent discussions with other students or resources outside of those 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