Homework 9

Due: Wednesday, November 8th, 2023 at 1:30pm on Canvas

Please submit both your RMarkdown file and your knitted HTML or PDF

**Part 1: Conceptual Questions**

1. Simmons, Nelson, and Simonsohn (2011) propose six requirements for authors. What are the six requirements? Pick three that you think are the most important/effective and explain why those three are particularly important.
2. What is “statistical influence” in the context of case analysis? What measures do we use to assess influence? What is the key difference between these measures?

**Part 2: Data analysis**

A public health organization wants to test the importance of staying active and healthy in the winter by examining links between cross-country skiing, vegetable consumption, and health, as measured by an independent doctor at the end of winter. Participants were 53 residents of the Madison, WI area who were recruited at the beginning of the winter and asked to record the number of times they went cross-country skiing over the course of the winter and their average daily vegetable consumption (in servings). At the end of the winter (i.e., the last week of March), a doctor conducted a health assessment of each participant and reported their health on a 0-10 scale, with 0 representing very poor health and 10 representing excellent health.

The researchers in the public health organization predict that cross-country skiing and vegetable consumption both contribute to overall health. They also expect to see little to no relationship between vegetable consumption and skiing behavior.

**Codebook for “hw\_09\_ski\_veg\_data.csv”**

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Values |
| sub\_id | Subject identification number | 1-53 |
| skiing | Number of times participant went cross-country skiing over the course of the winter | 0-93 |
| veggies | Number of servings of vegetables participant consumes on an average day | 0-16 |
| fitness | Level of health, as assessed by doctor, at the end of March | 0-10 |

1. Read in and inspect the data.
2. Fit a model testing the researchers’ hypothesis. Specifically, run a model testing whether skiing and vegetable consumption both uniquely explain variance in people’s health scores at the end of March (when included in the same model). Explain the result in a sentence in your R script.
3. You suspect there may be outliers present in the dataframe. First, identify which participants have high leverage. For each of these points, note the variable on which the point is extreme (i.e., why it has a large hat value). (Note: an observation can have high leverage (can be far away from the centroid) even if its scores on each of the predictors are not extreme).
4. Next, test for regression outliers. For each point you determine to be a regression outlier, report why this person is a regression outlier (in terms of the variables present in the dataframe).
5. Now examine the Cook’s D scores of the participants. Are the participants who have extreme Cook’s D scores surprising to you or not?
6. Produce the influence plot for this model. Which points have high influence as identified in this plot?
7. One datapoint has a relatively small Cook’s D value, but a relatively large hat value on one of the predictors. Who is this person, and why is their influence small despite their large leverage?
8. If you had to give the researchers advice about whether to remove any of the participants, what would you say?
9. Give an example of some rule you could have made regarding recruitment for the study that could have prevented the researchers from obtaining the kinds of outliers present in the data.
10. Regardless of your question 8 advice, remove all problematic participants (in terms of influence) and run your focal model again. What has changed? Why?
11. Write a 3-4 sentence results section explaining what you found in your case analysis (assuming you are presenting the analyses in which you decide to remove the problematic participants).
12. Make a publication-quality graph displaying the relationship between vegetables and health. Include two regression lines: one should show the relationship when controlling for skiing and one without skiing factored in. Does the line change much? Why or why not?
13. How many hours did this homework take you?