Lab 6 In-class Exercise

Friday, October 13th, 2017

**Part 1: Find the bologna!**

Imagine you are a reviewer on a paper. The authors mention removing 2 participants, but don’t give any rationale for doing so. You ask them about this issue. For each of the hypothetical responses below, rate how full of bologna they are on a 1-5 scale, where 1 is “vegetarian’s delight” and 5 is “bologna and bologna alone day,” then explain how you came to this conclusion (briefly). The authors haven’t provided you any preregistration documentation.

1. “We decided before running participants they survey should take at least 7 minutes to complete. One participant completed it in 5:15 and the other in 6:54.”
2. “Both these participants completed the survey about a minute faster than any other participants.”
3. “The data for both these participants were partially corrupted when a stressed grad student spilled coffee on the computers they were working on.”
4. “The participants didn’t fit in with our hypothesis, so we removed them.”
5. “The standard error decreased when we removed them, so we decided it was worth it.”
6. “The participants had scores on one of the predictor variables that were very uncharacteristic of the rest of the data. We rephrased our conclusions to be limited to the range of the predictor variable that was more typical.”

What’s one thing you could ask any of these research teams to do to make readers of the future publication less skeptical of their outlier decisions?

**Part 2: Data analysis**

A public health organization hopes to show the importance of staying active and healthy in the winter by establishing 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 his/her health on a 0-10 scale, with 0 representing very poor health.

The researchers 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 “Lab6Ex\_skivegdata.csv”**

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Values |
| SubID | 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. Explain the result in a sentence in your R script.
3. You suspect there may be outliers present in the dataframe. First, identify which points 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).
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). For this question, use a somewhat liberal definition of outlier.
5. Now examine the Cook’s D scores of the participants. Are the people that have extreme Cook’s D scores surprising to you or not?
6. Take a look at the influence plot for this model. Before you click the points of high influence, guess which point represents which of these problematic participants.
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 simple, 3-4 sentence results section explaining what you found in your case analysis (assuming you’re presenting the analyses in which you decide to remove the problem participants).
12. **BONUS**, for practice: make a publication-quality graph showing the effect of vegetables on health. Include two regression lines: one showing the effect when controlling for skiing and one without skiing factored in. Does the line change much? Why or why not?