Response to reviewers

Amaliah, Cook, Tanaka, Hyde, Tierney

Submission ID 217815520.R2

We thank the editor, associate editor and the reviewers for their feedback. Here is how we have made revisions to the manuscript to address each point.

Editor: The reviews are in general quite favorable and suggest that, subject to the minor revisions that the paper would be suitable for publication as part of the special issue on reproducibility and responsible workflow.

Reviewer 3: Thank you for this revision. My minor comments and corrections are all well addressed, and I particularly commend the authors for their changes to Section 3.4. It was a great idea to create a Shiny app showing the exploratory process that lead to your choice of threshold; this feels far more justifiable and reproducible than reporting the final number alone.

- Thank you!

Associate editor

While the data are interesting, the necessary details for a pedagogical application or case study are still largely missing. The manuscript reads more like a paper about the R package, something for a software journal. In my last review, I had missed that this was for a special issue, still I agree with the reviewer that the clear contributions to educators other than its use in class are plenty, but not fully described.

- Nothing to respond to here.

The main item left to be addressed is polish. This twice-revised manuscript still contains several places where authors are missing words, awkward phrasing, or grammatical errors.

- Corrected as detailed below.

The abstract promises a discussion of the process for cleaning data and exploratory analyses for longitudinal data. The latter is not provided, nor is a discussion about how that might be done in a class activity. This is a place for adding another clear contribution to educators. (Alternatively it can be deleted).

- The abstract does NOT promise an exploratory analysis for longitudinal data. It mentions that one of the purposes of this textbook data is to use it to teach exploratory analysis of longitudinal data. Similarly, it also states that the data is useful for teaching modeling. Both of these are beyond the scope of this paper. The mention is still important to keep because it makes the reader aware of why this data are useful and important for education. No change to abstract was made.

(page 10, line 18) “these” is errant

- It has been removed from the sentence.

(page 11, line 14) Providing one or two sentences about the use of pivot_longer() would help readers gain insight and motivate clicking the link to the code.

- Done.

(page 11, line 35) The sentence beginning “This information is provided” is missing at least one word.

- This sentence has been modified.

(page 12, line 49) There is still the error stating that education should only increase. Certainly, you mean non-decreasing.

- We have modified this sentence and change the word to be non-decreasing.

(page 13, line 24) I’m just checking that all Q... variables are correct. The formatting is inconsistent; e.g., Q3-8A versus Q1-3_A

- Yes it is correct, the formatting of all Q… variables uses the same format as the downloaded data from the database.

(page 13, line 36) The sentence beginning “Our target variables” is missing at least one word.

- We have modified this sentence.

(page 14, line 20) Why is 84 hours the cut-off? This seems to be an arbitrary choice.

- This corresponds to 12 hours per day, each day of the week, which is a reasonable upper limit on the expectation of the maximum amount of hours worked per week. It is quite a high upper bound, and higher than the IQR of 67. We have added clarification in the text.

(page 15, line 3) The sentence beginning “For stwork variable” is missing at least one word.

- We have modified this sentence.

(page 15, line 3) It might be helpful for readers if you refer to variables as what they are and put the column title in parentheses.

- Done.

(page 15, line 15) Providing one or two sentences about the use of join() would help readers gain insight into what the authors are doing here.

- We do not mention the `join()` function in the manuscript. It is true that we mentioned "the employment and demographic variables are then joined". However, in the current version, we have changed "join" to be "merged" so that it could reach people who are not familiar with `dplyr`.

(page 22, line 5) The default for rlm is huber weighting which never applies a weight of 0. You may have meant bisquare, but that change should be discussed.

- We are really not sure what you are requesting here. The section was re-written in the revision. We don't refer to huber weighting, and it makes no sense to refer to bisquare.

(page 28, figure 7) This is a helpful graph, but perhaps making the barplot "position dodge" and the density plots overlapping with transparency would sell your point better. This graphic makes the point, but the suggestions I made may produce more compelling evidence.

- We have done as you suggest for the density plots, but feel the back-to-back bar chart is better for the `hgc` comparison.

(page 29, line 21) The sentence "The highest grade completed has some confusion" should be rewritten.

- Done.

(page 29, line 37) Naming the function would be helpful for readers interested in using it.

- Done.

(page 31, line 9) Certainly you don't mean to say that you've created data "unsatisfactorily far from" the original.

- We have changed "unsatisfactorily" to "disappointingly".

(page 32, line 18-30) Here you describe some modeling (see first note). Re-framing this as a possible activity could be helpful to the reader. As it is written now it sounds like a summary (the section title) of what was presented – instead you can make it a path forward for the readers.

- This bullet list is a path forward. It is clearly not a summary of what was presented. We did not discuss teaching modeling. We simply used one type of model for pre-processing the data to handle unusual observations. No change made.