Review: The Journey from Wild to Textbook Data: A Case Study from the National Longitudinal Survey of Youth

This manuscript provides an interesting case study for demonstrating data cleaning. The motivation for doing this is clear and the questions that arise would make for valuable classroom discussions.

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.

I agree with the general sentiment among two of three reviewers that this paper is improved over the previous version, but still doesn't make *clear* contributions to educators other than its use in class. If this was (1) a general process for demonstrating the cleaning of longitudinal data, (2) a case study for cleaning data in the classroom, or (3) a case study for using the data in the classroom the value would be clear. In its current state, however, the manuscript still partly each of the three by describing the `R` package they've created and, as a result, doesn't fully address any of the three.

Below, I provide several comments about the paper. Some comments are verbatim from my original review because they have gone unaddressed; these are listed first and bolded. I haven't provided comments about grammatical or flow issues as the other reviewers have provided such feedback.

Specific Comments:

- **The author switches between gender (man/woman/etc) and sex (male/female/etc.) throughout the manuscript. Further, the explanation here covers important ideas in teaching how to use data on gender/sex and race, but it's just the beginning of ideas. It might be better to discuss the questions on the survey and discuss possible shortcomings, rather than make a somewhat empty gesture towards today's standard.** The authors still *incorrectly* refer to male/female as gender. While it is clear the authors are trying to be thoughtful in how they address important issues with data collection on demographics they are not correctly doing so (see note about race/ethnicity below). Certainly there are potential issues with binary sex variables, this can and should be discussed but care should go into this. Please review the literature surrounding the discussion of sex and gender in data (e.g., SAGER guidelines).
- **The discussion around highest grade completed is mechanical in what was done, but lacks substantive reasoning for those decisions and possible downstream consequences.** It is *still* unclear to me whether education is measured by year or by category (it's described both ways) and, in either case, the decision for one or the other isn't contrasted

or justified. For example `hcg` is described as a factor and `grade` is described as only increasing (numeric?) and surely the authors mean non-decreasing. If this is meant to be a case study for the classroom, this decision is rich ground for discussion.

- **I'm not sure this is accurate. The robustlmm package in R fits these models rather quickly with thousands of observations. What are the authors using? Does it run out of memory (requires more RAM) or does it take too much time?** The mixed effects models that robustlmm fits are identical to that of lmer. There is extensive documentation about the convergence issues (which are generally false positives). One may use the allFit() function to evaluate this.

- **The choice of 0.12 seems arbitrary at best. I don't see any justification for this approach (e.g., literature, simulations, etc). Why not use bi-square, instead of huber, and simply impute for observations with weight 0. While still arbitrary, it's decided by the bisquare model and isn't a guesstimate. Further, it would be sensible for any dataset and not just this one. (This would correspond to a smaller threshold, which may catch too many strange outcomes). In any case, it's hard to justify something so adhoc.** I agree that it is simpler not to use robustlmm (see note above about robustlmm), and that how to decide what to do with outlandish observations would be rich for class discussion. However, this isn't discussed as part of this data being a case study – if this is to be used for an educational purpose (beyond how the data were compiled) instructing this type of discussion and the possible avenues for exploration would be important. To teach this, we'd want to used a justified method for data imputation and outlier detection and note there may be other sensible ways (which I think this is) to do it.

Other Comments:

- The authors also make some problematic decisions about the race (Black or not-Black) and ethnicity (Hispanic or not-Hispanic) variables. They suggest three categories (non-Black/non-Hispanic; Black; Hispanic) and don't discuss the process of combining them, while recent standard is (non-Black/non-Hispanic; non-Black/Hispanic; Black/non-Hispanic; Black/Hispanic). It's possible that this selection was made for backwards compatibility, lack of certain demographics in the sample, or some other reason but it is not communicated. Please review the literature surrounding the discussion of race and ethnicity (e.g., Standards for the classification of federal data on race and ethnicity).

- (page 3-4, lines 34-20) The discussion about "the average and the individual" is underdeveloped. It seems to be tangled with a couple other ideas in these paragraphs, which muddles the point.

- (page 7, figure 1). Is this figure cut off or is it just at the end of the page?

- (page 9, line 4-6). "HRP1 1980 and HRP2 1980, contain the information about the job number up to 5..." They only contain job number one and two – in other areas of the manuscript you use a subscript i to make the point clear and this would be a good solution here. The code here goes off the page a bit, too.

- (page 10, tables 1-2) I believe these tables show what the authors see in their cleaned data compared to the NLSY numbers provided on page 15. Is this correct?