Response to reviewers

Submission ID 217815520.R1

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.

Major changes:

The editor says The paper falls squarely into the call for the special issue on "Reproducibility and responsible workflow". Some additional suggestions regarding framing are provided by the reviewers and the AE.

The only comments on framing appear to come from the associate editor: 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, does not fully address any of the three.

- It is not clear how to handle this. We interpret the comment from the Editor, that the paper is consistent with the call for the special issue. The AE would like it to be clearly one of three types. The paper does have of each of the three types, which we believe has value. There is some wording in the introduction ('An example of this is the wages data made public by Singer and Willett (2003) in their book, "Applied lon- gitudinal data analysis,' which can be used to teach generalized linear models, in addition to hierarchical, mixed effects, and multilevel models.") and we have added additional text in Section 5 Summary, on lessons learned, to help the reader understand the elements that might be used for their teaching.

Associate Editor: **ORIGINAL COMMENT: 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.** R1 COMMENT: 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).

- We have revised the language, made a note for educators on usage, and referred to the SAGER guidelines.

Associate Editor: **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.

- We have added more explanation of the two variables, and suggestions of where to use one and the other in classroom analyses.

Associate Editor: **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.

- Using `rlm` with the `nest` and `map` functions from `tidyr`, efficiently and simply fit these models. They can be easily explained in a classroom setting, and equip students generally with advanced data handling skills. Similar results would be obtained from using robustlmm, though. We have removed reference to memory and convergence issues. We have also removed a specific reference to threshold used. Having the particular number used is not important for the explanation of the approach and as we have seen from the review, is distracting for the reader. This section has been modified so that the approach is more understandable.

Associate Editor: **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.

- This is more of a comment than a request to make a change, but we agree that this is an area that could be a good classroom discussion. We have revised the section on handling extremes to make our explanation simpler, with links to the shiny app that we put together to assist in

the judgment and choice of threshold. We hope that this provides fodder for discussion. It could definitely be worked into a teaching exercise.

Associate Editor: 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).

- We have revised the language, referenced the Federal standard, and explained the reasoning for using a single variable.

Associate Editor: (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.

- We have revised the introduction to fix this.

Reviewer 1: The manuscript is much improved. I would like to see less description of how to tidy the data in Section 3 and more discussion of lessons learned (useful for statistics and data science educators and students) in Section 5.

- We have added additional bullet points on the benefit for educators in Section 5, in response to the overview comments by the Editor and AE (top of this response). We think it is best to keep content on cleaning and tidying the data, in a reproducible manner, to match the mission of the special issue.

Reviewer 3: One thing that could be made clearer throughout the paper – especially in Section 5 – is that the ultimate goal here is to get an equivalent dataset to the original that contains the additional years of data and better respects modern social justice norms. The authors use the terms "update", "refresh", and "re-create" throughout, and I think it is sometimes ambiguous whether this is a replication or update.

- We have changed all to 'refresh' as it is not correct if we say it as a replication because we did not use the exactly same variable and same result (esepcially ID included in the dropouts cohort.)

Reviewer 3: Here is what I think the strong takeaway in this work is: "Any longitudinal dataset used for education should have a sufficiently reproducible process to be updated with new data. The NLSY79 dataset is a great teaching tool that has become outdated, both because the dataset stops in 1994, and because the demographic data could be handled more delicately. What you have done here is offered a well-documented and reproducible process that expands

the dataset to modernity, and matches the original dataset decently well within its scope." This narrative doesn't require any major structural revision; only some wordsmithing to make sure this message of your contribution hits home throughout the paper.

- Thanks for this, exactly right! We have changed the wording accordingly, in various places in the paper.

Reviewer 3: Page 17: I remain slightly uncomfortable with the "hand-way-ness" of the modeling step. I don't think that "we tried this and it failed to converge" is a satisfactory explanation for dismissing a model that the authors themselves believe would be a better fit. Perhaps the discussion around robustlmm could be relegated to an appendix or supplement, with a bit more detail about why this model doesn't converge on the data. That way it would not distract from the details of the model that was actually used.

- Same point as made by the AE, and how we have addressed it is described above.

Reviewer 3: Page 18: Similarly, the justification of the 0.12 cutoff is much improved from the first version of this paper, but it still feels a bit strange. This sentence in particular throws me off: "That struck a balance between maintaining the natural variability of the wageswith minimizing implausible values." There are some big assumptions in that sentence to do with what is the "true" natural variablility of the wages and what is "truly" an implausible value. Is there any semi-objective measure we could use to justify the choice of 0.12 beyond the "eyeball test"? For example: maybe a plot of the variability of the imputed data for various threshold choices, showing that a low threshold leads to very high variance and a high threshold leads to low variance? Or: Can you make an argument from predictive power of the model, i.e., that a threshold around 0.12 trained on years 1979-2016 best predicts years 2017-18? Perhaps this is too much lift at this stage, but any amount of quantifiable justification would relieve a lot of the subjectivity around that 0.12 choice.

- See response above to the associate editor, about this concern.

Minor changes:

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

- Yes, it is cut off, we have made the size smaller, so it is not cut off.

Associate Editor: (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.

- We have modified the discussion around HRP to be more concise about what we have done with the variable. On page 9, we use the HRP to demonstrate the untidy form of the data. Hence, the sentence 'HRP1 1980 and HRP2 1980, contain the information about the job

number up to 5…' has been removed. In another part of the manuscript, we explain that if a respondent has multiple jobs, the `mean_hourly_wage` is computed as a weighted average of the hourly wage (`HRP`) for each job with the number of hours worked for each job as weights (provided that the information on the number of hours is available); if the number of hours worked for any job is missing, then the `mean_hourly_wage` is computed as a simple average of all available hourly wages. We hope that our modification makes the manuscript clearer.

Associate Editor: (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?

- The tables show the number by characteristics (age, sex, and race) in the input data (referring the term in the de Jonge and van der Loo's statistical value chain), i.e., the data we have extracted and tidied from the database. We are not saying it as a clean data as we have not treat the extreme values. We have modified the tables' caption to make this clearer.

Reviewer 2: On page 6, several times you use the word "plan," as in "We also plan to include additional variables" and "The plan is to create three datasets as follows." Has this been done in the paper, or are some of these future goals?

- These have been done in the paper so we have removed it.

Reviewer 2: The word "mutate" is used several times in the manuscript. While tidyverse uses are familiar with that as function name and "verb," I think a general audience is going to find it a bit jarring. Consider replacing with the word "create" or similar.

- We have changed the word 'mutate' to be 'create'.

Reviewer 2: On page 10, I don't know what this sentence means: "When this value was missing, 2012, 2014, 2016, and 2018, but available in the first form substituted accordingly."

- This sentence means that when the revised `hgc` variable were not reported, we only use the unrevised version of the `hgc` in that year. We thank the reviewer for this notice. We agree that this sentences is not clear, so we have modified it.

Reviewer 2: The word "that" is used a lot in this manuscript, and most instances could be removed. For example "For example, an article published in the Sydney Morning Herald argues that there is no average Australian" can be replace by "For example, an article published in the Sydney Morning Herald argues there is no average Australian." I suggest searching for the word "that" and removing any unnecessary instances.

- We have removed unnecessary instances.

Reviewer 2: Some extraneous punctuation marks are present

  - p1 in abstract: Both "wages textbook subset, have not" and "open source R package, called" do not need commas.
  - p2 "high school dropouts, from 1979-1994" does not need comma

    - p2 comma after “divergence of purpose” might be better replaced with dash.
    - p4 the sentence “Plot (C) shows the profile for an individual, with not such a high maximum wage but still indicates a problem: their wages are consistently low except for one year where they earned close to 1200/hour.” needs a few edits. I suggest “Plot (C) shows the profile for an individual with a maximum wage that is not so extreme but still indicates a problem: their wages are consistently low except for one year where they earned close to 1200/hour.”
    - p24 “predominately” does not need a hyphen.
- All of the extraneous punctuation marks mentioned have been removed.

Reviewer 2: p2 missing closing parenthesis after Stodel 2020 citation.

- We have added the missing closing parenthesis.

Reviewer 2: Punctuation should go inside quotation marks, not outside

    - p2 comma “Applied longitudinal data analysis”,
    - p4 comma “tame data”,
    - p6 comma “statistical value chain”,
    - p9 period “female”.
    - p13 period “number of weeks worked since the last interview”.
- Done.

Reviewer 2: Miscellaneous comments:

    - p6 “For example, use a single categorical race variable instead of the two binary race variables.” Perhaps missing a “we” before use?
    - p6 The sentence beginning “van der Loo and de Jonge” is jarring because of the lowercase name, particularly because it starts a paragraph and section. I recommend flipping the clauses to begin “In the context of official statistics, van der Loo and de Jonge…”
    - p18 “The year when the individual starting to work.” should perhaps be “The year when the individual started to work.”
    - p24 The sentence “On an individual level, one needs to know where I am in this data and does this data relate to me.” needs edits. I’m not sure people from the longitudinal study are likely to be looking at this data, so “where I am in this data” is not quite accurate. Rather, people might want to see how their characteristics relate to those in the dataset.
    - p26 Cooksey reference should have NLSY capitalized.
- All of the comments on the list above have been addressed.

Reviewer 1: Minor comments:

- P2, l46: missing right parenthesis.
- P9, l24: missing word(s) "this sometimes difficult the adjustment"
- P11, l43: "hours" worked
- P12, l32: unit in "weeks"
- P14, l38: Figure 3 shows...
- P21, l21: "the weeks worked since" ...
- P24, l21: The sentence "On an individual level, one needs to know where I am in this data and does this data relate to me" is awkward. Please revise.
- P34 l 38: Figure 5 (C) is mislabeled

- All of the comments on the list above have been addressed.

Reviewer 3: Page 2, line 42 – Parenthesis is not closed.

- We have added the missing closing parenthesis.

Reviewer 3: This comment from the first review is unaddressed: “Black” and “Hispanic” and “White” should be capitalized. (There is some dispute among style guides regarding “white”, but the other two are unambiguous.)

- We have put the capitalisation back in. Besides, we did not use the term ‘White’ in the dataset as we used the same term reported in the database, which is ‘Non-Black, Non-Hispanic’.

Reviewer 3: Page 7 lines 36-42:

- 1. The wages data of the whole NLSY79 cohort, including females.
- 2. A separate table of the demographic data of the whole NLSY79 cohort.
- 3. The high school dropouts’ wages data is closest to a refreshed version of Singer and Willett (2003)’s data.

1 and 2 are nouns and 3 is a sentence – I’m a little confused what this third dataset represents. I think you mean that you create a subset of (1) in the scope of the original data, to see how closely it replicates?

- Yes, you are correct, we have edited the point 3 to improve the clarity of this sentence.