


From: Dewi Lestari Amaliah dlamaleeah@gmail.com 
Subject: News from JSDSE editor
Date: 24 April 2022 at 11:32 am
To: Dianne Cook dicook@monash.edu, Emi Tanaka emi.tanaka@monash.edu

DA

Hi Di and Emi,

I hope you are doing well.

I just got the latest reviews from the editors of JSDSE from the last revision we submitted. Here, I forward the reviews.

Shall we set a meeting to discuss this? In the meantime, I am planning to address the reviews related to minor issues, typos, and copyediting.

Best regards,
Dewi

----- Forwarded message -----

From: **Journal of Statistics and Data Science Education**
<onbehalf@manuscriptcentral.com>
Date: Sat, Apr 23, 2022 at 2:35 PM
Subject: 217815520.R1 (Journal of Statistics and Data Science Education) A revise decision has been made on your submission
To: <dlamaleeah@gmail.com>
Cc: <nhorton@amherst.edu>

23-Apr-2022

Dear Dr Dewi Lestari Amaliah:

Your manuscript entitled "The Journey from Wild to Textbook Data: A Case Study from the National Longitudinal Survey of Youth", which you submitted to Journal of Statistics and Data Science Education, has been reviewed. You'll find the reviewers' comments below or as attachments at the bottom of the first revision screen.

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 reviews are in general very favorable and suggest that, subject to some additional revisions, your paper should be suitable for publication. Please consider these suggestions, and I look forward to receiving your revision.

Note that you do *not* need to submit a blinded version of your revision.

Please revise your paper accordingly. When you submit your revision, you should create a PDF file called "Reply to Reviews" (or similar) that contains your point-by-point responses to the reviewers' comments. Please be sure to blind your response file.

Upon receipt, your manuscript will be returned to the associate editor (and, possibly, the reviewers) who will determine if your revisions are satisfactory.

To submit a revision, go to <https://rp.tandfonline.com/submission/flow?submissionId=217815520.R1&step=1>. If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

If you have any questions or technical issues, please contact the journal's editorial office at jscott@stat.osu.edu.

We also ask that you upload your source files at this time. Please upload the following:

- An unblinded PDF of your final paper
- One zip file containing the TeX or Word file of your final, unblinded paper, including all figure files as separate files, and any additional files needed to compile your article. For example, .bbl and .bib files if you used BiBTeX (and we hope you did!), and any special macros.
- One zip file containing all files to be included as online supplements to your paper (appendices, datasets, code, etc). There is no need to include the TeX file(s) for online PDF files. Please use the Supplementary Materials for Review designation for this zip file. A README file containing a list of your online supplements is very helpful.
- PLEASE NOTE: Please do not use LaTeX cross-referencing between your main paper file and online supplement files.
- You will complete a license to publish form online after your paper has been placed into production. There is no need to upload or send a license now.

IMPORTANT: Your original files are available to you when you upload your revised files. Please delete any redundant files before completing the submission.

Upon submitting your revision, you will be asked if the manuscript has been submitted to the journal previously. Click `Yes' and then cut-and-paste the manuscript number from the email you received.

Because we are trying to facilitate timely publication of manuscripts submitted to Journal of Statistics and Data Science Education, your revised manuscript should be uploaded by 23-Apr-2023. If it is not possible for you to submit your revision by this date, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to Journal of Statistics and Data Science Education and I look forward to receiving your revision.

Sincerely,
Dr Nicholas Horton
Editor, Journal of Statistics and Data Science Education
nhorton@amherst.edu, nicholasjhorton@gmail.com

Comments from the Editors and Reviewers:

Reviewer: 2

Comments to the Author

Thanks to the authors for their attention to reviewer comments. I find this revision to be much stronger, and all my comments are minor issues or copyediting.

Minor issues:

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?

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.

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.”

Typos/copyediting:

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 replaced 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.

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 em-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

p2 missing closing parenthesis after Stodel 2020 citation.

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”.

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.

Reviewer: 1

Comments to the Author

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.

Minor comments:

P 2, l 46: missing right parenthesis.

P9 l24: missing word(s) "this sometimes difficult the adjustment"

10, is a missing word(s) and sometimes about the organization

P11, 143: "hours" worked

P12, 132: unit in "weeks"

P14, 138: Figure 3 shows...

P21, 121: "the weeks worked since" ...

P24, 121: 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 138: Figure 5 (C) is mislabeled

Reviewer: 3

Comments to the Author
General Comments

This revision is a much better fit to the audience of JDSSE. It now has a clear "personality" as a description of a process with tie-ins to academic work.

I particularly like the discussion of the "statistical value chain", which nicely mirrors the process of the data cleaning. I also think Section 4.3 does a good job summarizing the takeaway messages that an educator might take from this experience.

I very much appreciate the updated handling of the sex/gender variable and of studying race vs. racism, as well as the discussions throughout the paper of the focus on the individual story in data.

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.

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.

The above narrative doesn't require any major structural revision; only some wordsmithing to make sure this message of your contribution hits home throughout the paper.

Errata

Page 2, line 42 – Parenthesis is not closed.

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.)

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?

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.

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 wages with minimizing implausible values.” There are some big assumptions in that sentence to do with what is the “true” natural variability 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.

Associate Editor

Comments to the Author:

See attached.



UJSE-2021-015
2.pdf

