**Reviewers' comments for miWQS: Multiple Imputation Using Weighted Quantile Sum Regression**

We thank the reviewers for their invested time and feedback for this review. We have provided a response to each of the reviewer comments below.

-- Reviewer #1 –-

# > Review of miWQS: Multiple Imputation Using Weighted Quantile Sum Regression

>

> **\* Overall recommendation:**

> I recommend “Accept” with only one very minor suggested change.

We thank the reviewer for the positive feedback.

> **\* Overview:** Is the article important to R the community?

> Yes, I believe so. The paper states, “the miWQS package to analyze epidemiological studies with chemical exposures, but researchers may also apply the package to sociology, genomics, or other areas in public health and medicine.” I believe there are many other areas where the application of weighted quantile sum regression to interval censored data.

>

> This paper is really good, and I could not find many issues. Along with the paper is a R script and a R Notebook

that demonstrates the package miWQS in action.

We thank the reviewer for the positive feedback.

> **\* Article:**

> The paper clearly states the questions that the R package and the paper’s examples address. There are nice examples that demonstrate the use of the R package and address the questions.

We thank the reviewer for the positive feedback.

> I like the reproducible aspect of the Computational Details after the Conclusion.

We thank the reviewer for the positive feedback.

> The references do not have dates, but otherwise they look thorough.

We thank the reviewer for the positive feedback. The reference list has been updated to include years and journal names, following the journal’s reference style.

> I really like (and find it useful) the decision tree (figure 6) to help researchers in using the miWQS package.

We thank the reviewer for the positive feedback. We agree that the decision tree (Figure 6) is a helpful tool for researchers.

> *Is anything missing?*

> My only recommendation/ suggestion is concerned with p. 10 that states, “Printing the resulting object answers the research questions of interest.” I suggest explicitly restating the research question on p. 10 since they were stated “way back” on p. 4 .

As suggested by the reviewer, the research question is restated in example 2, on page 10 in the manuscript as follows. “Printing the resulting object answers the research questions of interest. The research aims are to determine the association of the mixture with cancer and to find the important chemicals (if the association exists).”

> **\* Package**:

> The miWQS package is described well. Along with the paper is a R script and a R Notebook that demonstrates the package miWQS in action.

We thank the reviewer for the positive feedback.

> The package documentation looks very thorough. I think the package functions are appropriate. The examples demonstrate how the package functions can compute the AIC to compare the models and how to do a Bayesian analysis, just to name a few. When I start using the package, I might come up with improvements for the functions, but I cannot think of any now. The examples lead me to think that the functionality is quite thorough.

We thank the reviewer for the positive feedback.

> **\* Recommendations**:

> Page 10 that states, “Printing the resulting object answers the research questions of interest.” I suggest explicitly restating the research question on p. 10 since they were stated “way back” on p. 4.

As stated above, the research question is restated in example 2, on page 10 in the manuscript as follows. “Printing the resulting object answers the research questions of interest. The research aims are to determine the association of the mixture with cancer and to find the important chemicals (if the association exists).”

-- Reviewer 2 –-

# > Overall recommendation

> I am on the edge of minor and major revision, but will lean more on the former. The methodology and the codes seems to be solid (although I was only vaguely familiar with quantile regression before reading the paper), while the text needs some work, but I think it is mostly a matter of rewriting things while thinking about someone who is not an expert of these methods (or application).

We agree with the reviewer that the manuscript would be improved with some reorganization and clarification for readers who are not experts in mixture analysis. When revising, we followed the reviewer’s specific points below.

# > Overview

> I can see the methodology being useful in some applications, and the package seems to provide extensions to some earlier packages, so even though I am not particularly familiar with the applications (or theory) of quantile regression, I can see the merit for the article being published. I cannot say whether the methods are up-to-date, also because the reference list does not include years?

We modified the references to include years while following the journal’s reference style.

# > Article

> The article contains pretty detailed description of the methodology used in the package, and how to use the package, but the organization of material could be improved. While the paper answers most obvious questions (what kind of data and models the package supports), these answers are sprinkled between sections and if one is not already familiar with these types of problems, it is pretty hard to follow the text. But this is mostly a matter of restructuring and perhaps adding some clarifications here and there.

As suggested by the reviewer, we have reorganized some of the material and added clarifications for readers less familiar with mixture analysis.

> More specific comments:

> Given that the package name suggest a very general modelling technique, the style of the article is occasionally somewhat difficult to follow due to the mixture of general description of the framework and the heavy use of wording related to the apparent application area of the authors (chemical components) which I know nothing of. For example the first and second paragraph of the introduction could be written in a way that it is more clear what kind of data and/or what kind of models package can be used for in general terms, and only then explain that these kind of things are important for example in chemical epidemiological studies etc.

As suggested by the reviewer, we reorganized the first two paragraphs of the introduction by forming three new paragraphs. First, we introduce the type of data for the package by using the first two sentences of the old first paragraph. Next, we inserted the old second paragraph, which discusses the different applied areas. Finally, we used the remaining sentences in the old first paragraph to introduce the miWQS package.

> For those not familiar with quantile regression, it takes quite a long time before the whole meaning of this is introduced. Instead, authors focus first on missing values, and when doing so leaves it bit unclear whether the missingness is related to covariates, outcome variables, or both.

As state above, we have reorganized the first two paragraphs in the introduction (on page 1). The second paragraph is inserted between the second and the third sentences of the first paragraph. The remaining sentences of the old first paragraph form a new revised paragraph.

> There seems to be extra colon on the reference to Figure on the fourth paragraph.

The extra colon is removed in the text.

> The introduction starts by saying that the package is used for modelling correlated continuous variables. This makes me wonder how one could then use miWQS for creating socioeconomic status index as authors suggest, as surely for example race (mentioned by authors) is not measured as continuous variable and there is certainly no meaningful ordering available?

This is clarified in the text when socioeconomic status was discussed on page 1: “Researchers may also create a socioeconomic status (SES) index that is generally composed of continuous correlated variables in the following domains: educational achievement, race, income, housing, and employment (Wheeler et al. 2017, Wheeler et al. 2019a). For example, race may be represented by percent of the population that is white. There are several examples of this in the literature (Wheeler et al. 2019b, Wheeler et al. 2020).”

> The section “Data Structure” starts in a way I wish introduction would have started. But authors are still unprecise in terms of what kind of assumptions the methodology has regarding the type of y, X, Z. Should they all be continuous, or can they be something else as well? y is apparently univariate? Can there be missingness/censoring in all of these or only in X? Related to this, is all missingness in the data assumed be due to the censoring, or can there be other types of missingness (such as MAR) as well? Some of these questions are answered later, but I would like to see these questions answered here when the actual data structure is presented, if not already earlier in the introduction.

We have modified the text to increase clarity, where these questions are answered in the first paragraph of “Data Structure” on pages 2-3.

> Isn’t censoring between zero and detection limit DL a case of left-censoring, not interval censoring?

While we understand the confusion, the censoring described in the vignette is technically interval censoring. Interval censoring occurs when a random variable occurs between two numbers, like zero and the detection limit, while left-censoring is below a certain number.

> In the documentation of the example dataset is it said that X’s were simulated from independent normal distribution, but Figure 2 shows strong correlations between the variables?

This is correct. We addressed the concern in a new section, entitled “Recommendations in using miWQS package”. The first paragraph of the new section is the first three sentences of the first paragraph of the old conclusions section. Next, two new paragraphs that describe the limitations in the univariate Bayesian and bootstrapping approaches have been added to provide guidance on imputation model selection. The two new paragraphs are:

“Additional covariates, ***Z*,** may be used in the bootstrap imputation and WQS models. However, the univariate Bayesian model does not include covariate information in imputing the BDL values. This makes any covariate confounders uncorrelated with the imputed concentrations BDL. Thereby, the WQS regression coefficients, such as the weights and overall mixture effect, may be biased towards zero (Forer, 2014; Little, 1992).

Another limitation of the univariate Bayesian and bootstrap imputation models is that the X’s are imputed independently while the actual X’s are correlated. This makes the correlations among the imputed BDL values of different components biased towards zero. One concern is that the mixture with independently imputed BDL values may introduce some bias in the health effect estimate if a large amount of BDL values is present. As an alternative, an imputation model could take advantage of the correlations to impute a potentially more precise estimate (Dong and Peng, 2013; Little, 1992). One such approach is the multivariate Bayesian regression imputation model (Hargarten, 2020), which we are evaluating in ongoing work.”

The last paragraph in the “Recommendations in using miWQS package” section consists of the remaining sentences of the old first paragraph in the conclusion section. The second paragraph of the old conclusions section formed the new conclusions section.

> In Example 1, the forumula uses notation beta\_0bˆ(T) but the parameters are explained as beta\_0, beta\_1 etc. I can infer that additional subscript b denotes bootstrap sample b but this should be written out explicitly, and how does one obtain say beta\_0 from these bootstrap samples?

We use the nonparametric bootstrap to approximate the sampling distributions of the WQS regression estimates by sampling with replacement. In this sample of the data, the WQS regression estimates are determined via maximizing a nonlinear likelihood. In WQS regression, the goal is to find the average of the bootstrapped weight estimates, instead of the average or standard errors of the other estimates. In the first paragraph of Example 1 (page 5), we clarified the purposes of the training and validation datasets by replacing the text

“Ideally, the data should be randomly split into a training dataset and validation dataset so that the WQS index is created in the training set and is tested in the validation set.”

with

“Ideally, the data should be randomly split into a training dataset and validation dataset. While the training set is used to create the WQS index, the validation dataset is used to assess the association of the weighted index with the outcome.”

Additionally, we replaced the first sentence of the second paragraph in Example 1 (page 5) with

“In the training dataset, the weights are estimated from B bootstrapped samples of size to form the weighted index.”

> The constrain on beta\_1 seems weird. Is it needed for identifiability purposes? What if the researcher does not know the direction of the effect, or if the hypothesis is that beta\_1 = 0?

As suggested by the reviewer, we clarified the purpose of the constraint in the manuscript. The confusion arose as an overall mixture effect was in the training and another in the validation datasets in WQS regression. As indicated above, we clarified the purposes of the training and validation datasets with this sentence: “Ideally, the data should be randomly split into a training dataset and validation dataset. While the training set is used to create the WQS index, the validation dataset is used to assess the association of the weighted index with the outcome.” While a constraint is placed on in the training dataset, no constraint is placed on in the validation dataset. The hypothesis test is a one-sided test, usually in a positive direction for increased risk. The following paragraph is added after “The final weight estimate is calculated as an average of the bootstrap estimates for the *jth* chemical.”

“A constraint is placed on to allow for interpretation of the index (Carrico et al., 2014). Often, exploratory single-chemical analyses, shown in Appendix 1, show that some components in the mixture have a negative association with the outcome, while others have a positive association. In environmental risk analysis, researchers are often interested in a positive association between the mixture of components and an adverse health outcome. However, if a researcher hypothesizes that the overall mixture is protective of the outcome, the constraint > 0 should be switched to < 0.”

> I have some reservations regarding the bootstrap-imputation of the censored data. First of all, covariates z are used to predict the missing values in x but the original model formulation does not assume any dependence between these two types of variables? And in case of multiple variables with missing values, the imputation is done independently for each variable, which is of course an easy solution but I wonder what kind of errors this introduces if the observed X’s (above the DL) were strongly correlated?

We have discussed the discrepancy between the use of the independent normal distribution and the correlation in the index components in the new section, “Recommendations in using miWQS package.” The use of the independent normal distribution in the imputation is common to both the bootstrap and univariate Bayesian imputation models. This discrepancy led us to conduct a simulation study to assess the performance of the approaches. Our result showed that these univariate imputation models did well in identifying the important components and overall model fit when the proportion of BDL values was under 50%. These approaches had the best sensitivity, specificity, MSE, and model fit (Hargarten & Wheeler, 2020a).

Reference: P. M. Hargarten and D. C. Wheeler. Accounting for the Uncertainty Due to Chemicals Below the Detection Limit in Mixture Analysis. Environmental Research, 186:109466, July 2020a. ISSN 00139351. doi: 10.1016/j.envres.2020.109466. URL https://linkinghub.elsevier.com/retrieve/pii/S0013935120303595.

> Jeffery’s prior should be Jeffrey’s prior.

This typo is fixed in the text.

> In the conclusions section, there are some recommendations on what type of approach should be used in the presence of missing data. But, given that the handling of censored data is the main contribution of the package, I’d like to see this expanded more perhaps already in the previous Sections (and justified more than just citing some papers).

We have expanded upon the comparison of imputation methods in censoring data in the new section, “Recommendations in using miWQS package.” We discuss the limitations of the bootstrap and univariate Bayesian imputation models in that they do not include other component information in their imputation. We also discuss that covariate information is not yet implemented in the univariate Bayesian approach.

> While packages used by the miWQS are properly cited, there is no reference for R itself. The reference list is also missing the years and the journal names.

According to the journal’s instructions, we did not include a reference for R (Instructions for Authors, p. 7). The reference list has been updated to include years and journal names, while following the journal’s reference style.

# > Package

> Package is already at CRAN and seems to pass all automatic checks. There is also a link to a Github repo for issue tracking etc, but the repository seems to be somewhat weirdly structured and it is not up-to-date with the CRAN version, so it looks bit abandoned.

The miWQS package in the vignette is version 0.4.0. If this article accepted, the package will be updated on CRAN to version 0.4.0 so that the manuscript can be posted on CRAN as a reference for R users. Due to the resubmission policies on CRAN, we can only update the package once during this process, as listed in submission guidelines:

<https://cran.r-project.org/web/packages/policies.html#Re_002dsubmission>

> The code seems to be well documented and the source code is readable. There warnings when you the optimization does not converge etc, and there are some checks for function argument values, but if you set the number of bootstrap samples in estimate.wqs to zero you get out of bounds error (negative values are issued a proper error).

The error with zero bootstraps is fixed in miWQS version 0.4.0.

> I see no obvious issues in terms of efficiency (of course the Gibbs sampling could be implemented with C++/Rcpp etc but no R inferno stuff going on).

We agree with the reviewer that there are no issues with efficiency.

> At least in estimate.wqs and analyze.individually there should be a “. . . ” argument or other way to pass additional arguments to “glm2” and “solnp” functions.

For both the **glm2** and **solnp** functions, we prefer to add these options on a case-by-case basis, rather than in a general sense. The additional arguments available to pass in **solnp** is the “control” argument. As the control argument is usually created for sophisticated users, we are weary to allow general programmers access to change these settings. In **glm2** package, there are options to pass many arguments, including data options (i.e., weights, subset, and na.action), starting values and control features (i.e., start, etastart, and mustart, control, and singular.ok), and additional objects to return (i.e., model, x, y, contrasts). The data option arguments do not apply in WQS regression for two reasons. The `estimate.wqs()` function has already taken care of options subsetting and missing data. Additionally, weights doesn’t make sense in the context of weighted quantile sum regression. Arguments like changing the start values or control features in **glm2** requires sophisticated knowledge, so we are weary to allow general programmers access to change these options. Lastly, we are unsure what to pass or save for each model because each bootstrapped model and for the overall model uses the **glm2** package.