Review of "Technical Note: Effects of Uncertainties and Number of Data Points on Line Fitting – a Case Study on New Particle Formation", revised, by Mikkonen et al.

This is a review of a revised paper comparing various linear regression methods that account for errors in the x- and y-variables, which are also compared with ordinary least squares (OLS). The paper introduces the problem of linear regression with errors in both variables, and describes the method they used for generating synthetic data that is meant to represent data collected in new particle formation (NPF) studies (similar means, distributions, and noise levels). The results of the various fits are compared using various sized data sets, various noise levels, and data with extreme outliers included. Conclusions are drawn and recommendations made as to the preferred regression method(s) to use in particular situations encountered with NPF data.

General Comments.

This is a revised paper, changed according to comments made by two reviewers. The recommendations of the reviewers were acknowledged, but not all the suggested changes were implemented. This reviewer has no problem with this, but clear and justifiable reasons for not making recommended changes should be stated. This was not always the case in the author's response to the reviewers. This is discussed below.

In making the changes that were suggested, sometimes the English was not carefully checked. The original version of the paper had some minor issues with English. The revised version has many more problems. This reviewer suggests a careful review of the English, perhaps with the help of a native English speaker.

The heart of the paper depends on the analysis of synthetic data and the methods for its generation and simulation of noise. It is clearly explained what was done, but it is not always clear why. This reviewer suggests a detailed re-write of section 3 to make clear the decisions made on the approaches used to generate the data. Also, information should be added to the supplement to show the impacts of the data generation process on the final data. My concerns in this regard are addressed below. Also included below are other issues with the revised paper.

In general, the paper as written is quite short. It can easily be expanded to include additional important information that would make it more useful to researchers using various linear regression methods. This reviewer recommends that the authors feel free to provide any important information that would be useful to potential readers of their paper.

Synthetic Data Generation

The generation of synthetic data is routinely used to test data analysis methods. This is a suitable approach and can be useful in uncovering errors. If the data are not produced properly, bias and errors can result even when the analysis approach is correct. Consider the simple case of a series of measurements of a quantity near the detection limit. The calculation of the H_2SO_4 concentrations as described on page 5 is a good example. One thousand data points are randomly selected from a log-normal distribution with mean of 2×10^6 molecules-cm^{-3}. The standard deviation is 2.4×10^6. This generates concentrations from the mid-10^4 to the low-10^7 range. The recovered mean ranges from about 1.9 to 2.1 x 10^6, and the standard deviation from about 2.0 to 2.8 x 10^6. These values show a range because of the finite size of the data set and the random selection of data from the distribution. Next noise is added that has two components: a constant factor (representing the baseline noise) and a factor proportional to the value (representing signal-carried noise). The noise is selected from normal distributions with means of 4.0×10^5 for the constant part and 0.3 times the value for the proportional part. The noise ranges from about -10^7 to $+10^7$. Note that the errors are about the size of the largest values generated from the log-normal distribution. This is not realistic, since data below the detection limit are typically filtered out, and data considered valid are well above the baseline noise (typically 3 standard deviations). Indeed, there are several points in Figure 1 of the paper for

H_2SO_4 values less than 10^5 cm^{-3}. Note that the $J_{1.7}$ values do not follow the trend of the fit lines for these low H_2SO_4 values. This is because there are nucleation rates calculated that are negative and are eliminated from the log-log plot that would balance these values. In any case, this generates data whose lower values are negative (about 40-60 values out of 1000) and are undefined in logarithmic space. This creates a small bias with the mean of non-negative data about 3% to 10% larger than that used to generate the data. The standard deviation is 1% to 15% larger. The effect is even greater for the calculated nucleation rates (about 270 to 300 negative values) because of the 3.3 multiplier and the negative intercept. The mean for the nucleation rate is about 1% to 70% larger than the noise-free data set, while the standard deviation is 0.1 to 1.9 times that in the noise-free data. These biases affect the quality of the linear fits.

This reviewer suggests a different procedure for generating synthetic data. Rather than sample from a normal distribution, suggest generating evenly spaced data from some minimum value above the detection limit to some typical largest value observed. This would produce data that evenly covers the range of expected values, rather than a clumping data near the mean value of a distribution. In looking at the paper by Kurten (ACP, 2019), it appears that room temperature H_2SO_4 concentrations in CLOUD NPF experiments range from 3 x 10^7 to 10^9 cm^{-3}. Based on the noise levels assigned in this paper (Mikkonen et al.), this appears to be reasonable with a range from the detection limit (about 3 times the noise of small values) to the largest value measured. In tests, this reviewer configured the data (1000 points) evenly spaced in the logarithmic domain, with natural logarithms ranging from 16 to 21. This produced a mean H_2SO_4 concentration of about 2.6 x 10^8 cm^{-3} and a standard deviation of about 3.2 x 10^8 cm^{-3}. This results in 0 to 1 values that are negative out of 1000. Calculations of the nucleation rate only produces about 20 to 30 negative numbers. This leads to data that are much more reasonable (see Figure R1) in that the data are scattered evenly without unbalanced "tails", particularly at low H_2SO_4 values. It is worth noting that the nucleation rates are very large at 10^9 cm^{-3} H_2SO_4, but these are the values obtained from the equation $\log_{10}(J_{1.7}) = 3.3*\log_{10}(H_2SO_4)-23$ provided in the paper.

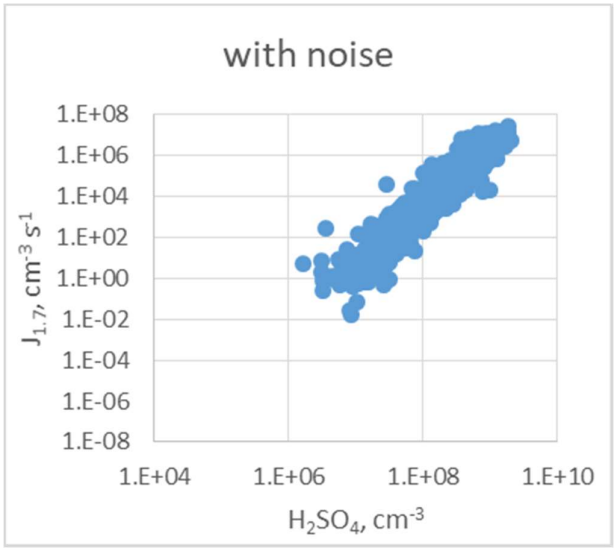

Figure R1. Synthetic data generated evenly spaced in logarithm space. Includes noise added to H_2SO_4 and $J_{1.7}$ values as described in Mikkonen et al., revised. OLS fit yields a slope of about 3.1, an intercept of about -21, and r^2 of about 0.9.

Note that least squares routines have problems if the dynamic range is not large compared to the noise in the data. In the above example, if the data range is significant narrowed, the data show no obvious trend, and OLS gives a distinctly flatter slope (Figure R2).

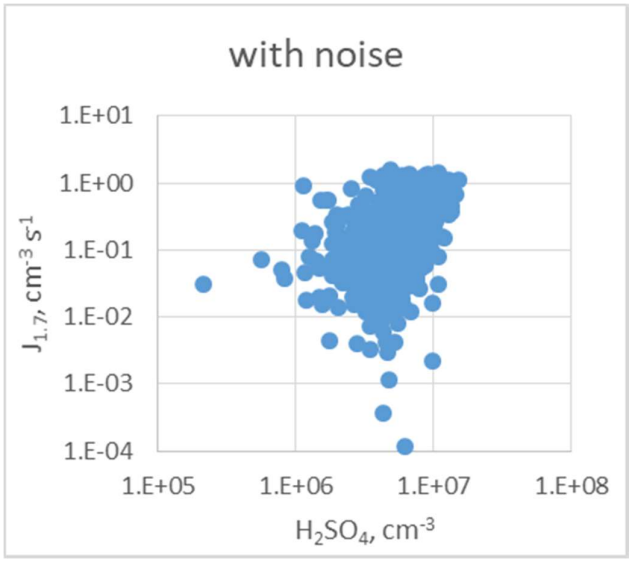

Figure R2. Synthetic data generated as in Figure R1, but with minimum and maximum $\ln(H_2SO_4)$ =15 and 16, respectively. OLS fit yields slopes of about 1.2, intercepts of about -9, and r^2 of 0.2. This appears to be caused by data covering a narrow dynamic range compared to the noise level.

In the generation of synthetic data, there is a random factor associated with the noise estimation (selection from normal distributions). This means that there is variability to the data and thus to the fits. It would be beneficial to run several (perhaps tens or even hundreds) 1000 data-point sets and give information on the variability of the fit parameters (minimum, maximum, mean, standard deviation of the slopes and intercepts of the various methods). This would allow other researchers to have better understanding of the range of values that can be expected for the different methods.

Since the authors likely have access to significant NPF data from CLOUD and Hyytiala, it is possible they could perform fits using what they believe to be the appropriate fitting method and compare them to the literature values. This would call attention to using the correct procedure, and would provide a database of corrected data for use by the aerosol community. This reviewer does not believe this is beyond the scope of this paper. It would involve a paragraph of introduction to the data, a description of the fit method selected, and a table of results.

Weighting

One of the advantages of many of the non-OLS fitting methods is the ability to weight data based on some factor, typically the inverse of the uncertainty. This can minimize the effects of outliers and cause the fits to depend more on data that is more certain. This reviewer suggests a discussion of weighting be included, and its impact on fitting be demonstrated, including comparisons of fits with and without weighting, and the impact of different weighting approaches on the fit results.

Addressing comments from original paper

This reviewer felt that several comments from the original paper were not properly addressed. This is not acceptable. The role of the reviewer is to make sure the presentation is scientifically robust and the paper justifies the approaches taken. It is important the sufficient information be given that the research can be reproduced. There were several comments that this reviewer does not believe were adequately addressed in the author's response to the reviews, and in changes to the paper.

1. Comment about more information on iterative methods.
   The response indicated that the information requested was in the Supplement. This is mostly not the case. While the functions to be minimized are given for many of the methods, the

criteria for convergence are not indicated. Also, no information is given on the York method, but only reference to the York, 2004 paper. Iterative methods rely on convergence criteria that indicate if sequential iterations vary by less than some fractional value, the procedure is halted. If incorrect convergence criteria are used, the procedure could be halted prematurely They also are prescribed a maximum number of iterations and initial guesses for the parameters. It is possible that if the maximum number is reached (because too small a value was selected) before convergence is reached, then the fit values will be incorrect. Poorly selected initial values can also inhibit convergence. This reviewer suggests using a large maximum number of iterations and repeating the fits with convergence criteria that are gradually tightened to see if the fits are changed substantially. In any case, the paper and/or the supplement need to address clearly the issue of using iterative methods, and indicate the convergence criteria and number of iterations prescribed for each one. This reviewer does not feel that keeping the convergence criteria at the defaults by the software programmers is prudent, sensible, or sufficient to address the concerns. Sensitivity tests must be done. Also, the statement about the York method (a custom python implementation) in particular calls for tests with known data sets to ensure it is functioning as it should. More detail about the York method also needs to be added to the Supplement section.

The equation given for deriving the fit parameters for the BLS (Francq and Govaerts, 2014) does not agree precisely with their Equation 24. Suggest checking Lisy et al., 1990 and other related papers to make sure equation is correct.

2. In the author's comments (page 8) related to the original manuscript page 4, it is stated that the "true" values are known because of the synthetic generation procedure used. While this is true before the noise was added to each variable, it is not necessarily true afterwards. This is because of the issues discussed above in which negative data are eliminated when conversion to logarithm space, which can potentially create biases. This is the argument for using other data in which the slope and intercept are known and established. This reviewers insists that at least one other data set be tested with each of the methods and compared with the known, exact fit parameters.

Other comments.

Suggest defining terms that might not be familiar to atmospheric scientists, such as "homoscedasticity" and "heteroscedasticity". The terms "estimators" and "predictors" are also used without definition, as is "a posteriori".

Manuscript specific comments

Page 1, line 10. Suggest eliminating "on a scatterplot" to read "Fitting a line of two measured…"

Page 1, line 10. Suggest removing "as", removing "considered" and changing "simplest" to "most common" to read "…variables is one of the most common statistical…"

Page 1, line 21. Suggest "Atmospheric measurements always come with some measurement error."

Page 1, line 23. Suggest rewording and/or adding text to clarify what is meant by "ill-formulated".

Page 1, line 25. Suggest "Regression models can be linear or non-linear the selection of which depends on the data being analyzed."

Page 1, lines 26-27. Suggest "…that the independent variable of the model has been measured without error and the model accounts only for…"

Page 1, line 29. It is not clear why OLS should be asymptotic, since it is not an iterative method. OLS can be close to the correct parameter value if the noise in the independent variable is small. Suggest rewording this sentence.

Page 2, line 1. Suggest changing the sentence "Measurement error needs to be taken into account" to something that indicates that methods that account for measurement error in the independent variable need to be utilized. In the next sentence that starts "Thus, we chose such…", suggest changing

to something that says that test data were developed that included significant uncertainties in both the independent and dependent variables.

Page 2, line 4. Suggest removing "as" to read "…is typically assume to be…"

Page 2, line 5. Suggest giving the units for J and H_2SO_4.

Page 2, line 7. Suggest "…to estimate the effects of new particle…"

Page 2, line 11. Suggest adding information to describe what is meant by "unconstrained" in this context.

Page 2, lines 13-14. Suggest "…of this inconsistency is very likely due to use of different fitting methods." Also, "…has been acknowledged previously in…"

Page 2, line 26. EIV methods simply mean that errors in both variables are accounted for. Suggest a statement that says this.

Page 2, line 33. This sentence implies that ORDPACK is unique in accounting for point by point variance and covariance, but other methods also have the capability (including York). Suggest rewording.

Page 3, line 7. It is misleading to say that in linear regression bias cannot be taken into account. Indeed, it is important to minimize bias through careful and regular calibrations and zeros. Analysis of these data can reveal information about baseline noise levels and signal carried noise. At a minimum, an upper limit to the amount of bias can be estimated, although obviously not known with absolute certainty. Suggest rewording.

Page 3, line 11. "…of the variable being measured."

Page 3, line 13. Suggest rewording "Natural error is that the true connection…" to something like "Natural error is the variability caused by natural or physical phenomenon"

Page 3, line 16. Suggest "…when interpreting fits."

Page 3, lines 17-18. Suggest "…in some cases this has to be taken…"

Page 3, line 19. Suggest "…independent of each other…"

Page 3, line 23. Suggest "…room temperature in the measurement space and atmospheric pressure may affect the performance of instrumentation…"

Page 3, lines 28-29. In my previous review, the point was that the sentence about the CERN NPF data is not necessary and should be eliminated.

Page 4, lines 1-2. Suggest rewording the sentence starting with "Existing data…" to something like "The existing data on NPF includes what are believed to be the most important routes that involve sulfuric acid, ammonia and water vapor…"

Page 4, line 6. Rather than "These variables…" suggest "The relationships between precursor gas-phase concentrations and particle formation rates were chosen for study because…". Suggest eliminating "…which makes them good illustrative variables for this study."

Page 4, lines 9-10. Suggest rewording to something like "…in the analyses, which casts doubt that errors have been treated properly."

Page 4, line 11. Suggest "…to have a linear relationship, but in order to raise awareness in the aerosol research community, in this study we relate our analysis to the important problem of understanding new particle formation."

Page 4, lines 24-25. Suggest "…how they account for measurements errors. The minimizing criteria for all methods are given in supplement S1, but here we give the basic principles of the methods."

Page 4, line 28. Suggest "…the error variances, λ_{xy}, of the variables…"

Page 4, line 29. Suggest "The approach of PCA is…"

Page 4, line 33. The phrase "linear scale uncertainties in logarithmic scale regression" needs explanation. Suggest adding a line or two to clarify, and also to explain why the York method cannot account for this.

Page 5, line 3. Suggest "…in both variables, and thus is a more advanced method than DR…". Note than many of these methods allow entering point by point uncertainties and thus can handle heteroscedasticity.

Page 5, lines 4-5. This sentence doesn't make sense. One method accounts for measurement variance, while the other methods require estimates of measurement errors. Isn't this the same thing. Suggest rewording.

Page 5, line 5. Suggest "Though for Bayes…"

Page 5, line 6. Suggest removing comma "…be applied with both errors given…"

Page 5, line 8. Suggest "…was calculated with the package…and BLS with the package…". Consider putting the software package names in quotes or underlining to make it clear they are special words.

Page 5, line 14. Upper case for the independent and dependent variables, but that is different than equation 1 and the supplement. Suggest making consistent throughout.

Page 5, line 24. When this reviewer performed statistics on the noise free values calculated for $J_{1.7}$, there was considerable variability from one 1000 point data set to the next. Giving the mean and standard deviation for one data set is rather meaningless. Either perform statistics on many data sets and given means of the individual means, or eliminate this sentence.

Page 5, line 26. The word "true" is used here to mean the simulated measured values (data with errors included) whereas elsewhere "true" is used to mean the input values (without errors). Suggest a different word than "true" here, and use care to be consistent throughout the paper.

Page 5, line 28. The second sigma should be changed to "$\sigma_{rel,y}$".

Page 5, line 32. Outliers are also generated on the low tail of the distribution. These should be eliminated as well.

Page 6, lines 4-6. In a list, suggest using "first…second…third" or "firstly…secondly…thirdly". In other words, be consistent.

Page 6, line 8. Suggest "noise-free slope" instead "true slope" for reasons mentioned above.

Page 6, lines 14-15. Suggest "…and performed fits with each method on all of these datasets."

Page 6, lines 15-16. Suggest "…to Figure 1 marked in black."

Page 6, lines 16-17. Suggest rewording the sentence that begins "It shows that when…" since the first half and the second half say the same thing, just reversed.

Page 6, line 17. Suggest "…which overestimates the slope…"

Page 6, line 19. Suggest "…are not changing significantly."

Page 6, lines 23-24. Suggest "…to their characteristic levels…"

Page 6, line 25. This recommendation is misleading because it depends on the noise (error) level of the data. For the conditions of this study, it may be true, but could be very different for data with more or less noise. Suggest rewording.

Page 7, line 7. Suggest '…on data, such as "How is Y related to X?".'

Page 7, lines 9-10. Suggest rewording or eliminating "…because methods measure slightly different things about the data." It is not clear what is meant by this sentence.

Page 7, line 9. The conclusion that including of error in the analysis will never lead to a more biased estimator was not discussed in the paper, nor is it justified by the material presented. Suggest either adding a discussion of this point with data that proves it, or eliminate the sentence.

Page 7, lines 20-21. It is not true that all fit methods highly uncertain with small numbers of points. Some methods are exact with small numbers of points (e.g. York method and Pearson's data with York's weights). This needs to be reworded or eliminated. Also, true of related statements in lines 31-32.

Page 7, line 29. A new symbol was introduced without definition "rE". Suggest either eliminating or defining. This could have been defined and used earlier in the paper.

Page 7, line 32. Suggest "Our study showed BLS to be the most robust with…". Maybe true, but depends on the uncertainty of the data.

Page 8, lines 1-2. It is not clear how error distributions are used in the Bayes EIV method. Indeed, there are symbols used in the supplement describing the Bayes method that are not defined. This sentence and the related section in the supplement need to be reworded.

Supplement specific comments.

Throughout the supplement, define variables used. Also, equations need to be numbered so they can be referred to.

Page 1, line 3. Suggest and introductory paragraph that describes what is to follow for each of the methods. Suggest headers to separate the discussion of the different methods. Suggest some more discussion of the synthetic data generation such as showing probability density functions for each variable graphically and perhaps other useful information that would be helpful to the reader.

Page 1, line 15. The sentence that starts "ODR takes into account…" does not make sense. This needs rewording to make clear. How can the errors be accounted for but not the variances?

Page 1, line 18. There is no error in the Y-axis, only error in the Y-data. Suggest making this change here and throughout the paper.

Page 2, BLS section. The first equation looks exactly like OLS with weights. I'm sure there is more to it than that, so more explanation is needed. As stated before, the second equation doesn't agree exactly with that in the Francq and Govaerts, 2014 paper.

Page 2, line 9. Suggest "A second bivariate regression method that was used in this study is an…"

Page 2, PCA section. This again looks just like OLS. Suggest adding more text to explain how it is different.

Page 2, line 26. Suggest "…and are treated as unknowns."

Page 3, line 1. Suggest "The Stan tool solved regression problems using 1000 iterations, and it provided…". Also suggest "In our analyses, we used the maximum a posteriori estimates for β and β provided by the software tool." What is the iterative formula used in this method? How do you know 1000 iterations is enough for full convergence? Is there a convergence criterion that indicates the software can finish? How do you know the criterion is appropriate? (These questions are related to general comments made earlier, but they apply to each of these methods.)