Review of “Technical Note: Effects of Uncertainties and Number of Data points on Inference from Data – a Case Study on New Particle Formation” by Mikkonen et al.

This paper begins by discussion of using regression to infer aspects of observed data and goes on to describe the issues related to new particle formation rates. The heart of the paper involves generation of data purported to represent the logarithmic relationship between new particle formation rates and sulphuric acid concentrations. Several datasets are produced varying the amount of uncertainty, the sample size, and the number of outliers using seven regression procedures. From the regression fits, the paper makes recommendations as to when various procedures are appropriate.

This reviewer found the paper interesting and relevant for studies of atmospheric measurements, but in some cases the detail was not enough to assess the value of the results or the recommendations presented. The paper needs significant work before it is ready for publication. The authors should review the recommendations of the reviewers and make the needed changes. Perhaps the revised paper will be suitable for publication.

General comments.

One of the key points of the paper was the inclusion of accurate estimates of errors in linear regression. This reviewer found the discussion of errors significantly lacking, and indeed including some incorrect statements. Within a measurement, there are two types of error: random and systematic. Random errors can come from natural atmospheric fluctuations and instrument noise. Systematic errors can come from errors in calibration and loss of analyte in the inlet. This reviewer has never heard the term (nor could I find reference to) “natural error”. One of the papers referenced (Carroll and Ruppert, 1996) also discusses “equation error”, which refers to the errors associated with using an inappropriate form of a fitting equation. The paper needs a much more thorough description of errors, including introducing the symbols used later in the paper to describe errors.

The paper states on page 3, line 12 that the data used in this study are new particle formation rates and sulphuric acid concentrations. In fact, the data are simply calculations of two variables related by a linear relationship with noise added to represent random and systematic uncertainties (as done in other previous papers on linear regression). The data could represent any relation that is expected to be linear. The paper does not address nor answer any of the issues related to measurement or calculation of new particle formation rates except to say that one needs proper error estimates to perform regression on observed data, and that there are significant differences found depending on how data is handled. The reviewer finds this attempt to connect a linear regression paper to new particle formation without actually directly addressing the issue misleading. One solution would be to change the title, eliminating the part of about new particle formation, and to simply present new particle formation as one example of where error estimates are important for linear regression. With the current title, the paper needs much more emphasis on the issues related to determining new particle formation using measurements and regression procedures.

Several regression methods are used in the analysis, but the information about their use is superficial. For example, many of the methods are iterative. If proper convergence criteria are not set, then the results obtained are not useful. It is important to state the convergence criteria for each iterative method and state how it was determined that convergence was reached. For other methods, if there are adjustable parameters, these should also be discussed. Also, the software or program used for each of the methods should be given. If they are programs written in-house, it might be appropriate to make them available to the reader.

Specific Comments

It should be mentioned, perhaps in the introduction, that linear regression is appropriate when there are two measures of the same quantity (for example, by two different instruments) or when there are two measures that are related by a physical law (for example, the dependence of the logarithm of a rate coefficient on inverse temperature).

Page 1, line 20. Suggest changing “comes” to “come” since strictly speaking “data” is plural (although often used singular).

Page 1, line 22. Did not understand the “making inferences in some more general context than was directly studied”. Suggest rewording or adding more information.

Page 1, line 23. Suggest “…the bias in the analysis method…”. Sentence needs period.

Page 1, line 29. After “…coefficients are underestimated…” suggest adding a reference.

Page 1, line 29-30. Suggest “Measurement error needs to be taken into account, particularly when errors are large.” Suggest removing “Thus, we chose such parameters as our test variables in this study.” Suggest replacing it with “To demonstrate this point, we show the effects of large errors on linear regression in this study.”

Page 2, line 1. Suggest “…known to strongly affect the formation…”.

Page 2, line 3. Suggest “...between *J* and H_2SO_4 is typically assumed to be of the form: …”.

Page 2, line 6. Suggest “…formation on global aerosol amounts and characteristics. Theoretically in homogeneous nucleation, the slope of this relationship is related to the number of sulphuric acid molecules in the nucleating critical cluster, based on the…”.

Page 2, line 9. Suggest “…results have shown discrepancies in the expected *J* vs. H_2SO_4 dependence.”

Page 2, line 9-11. Suggest “Analysing data from Hyytiälä in 2003, Kuang et al. (2008) used an unconstrained least squares method and obtained β=1.99 for the slope, wheras Sihto et al. (2006) reported a value of 1.16 using OLS from the same field campaign.”

Page 2, line 12. Suggest “…different time windows, but a significant proportion of this…”.

Page 2, line 14. Suggest “…fitting method as presented in York…”

Page 2, line 15-16. Suggest “…of the methods that do not need to know the errors in advance, but instead made use of estimated variances.”

Page 2, line 16. Suggest “Here, we present appropriate tools for using that approach.”

Page 2, line 17. Suggest “…have been made to present methods accounting for errors in predictor variables for regression-type analysis, going back to Deming (1943).”

Page 2, line 19. Suggest “…due to its simplicity and common availability in frequently used software.”

Page 2, line 20. Suggest “…methodological papers utilizing similar…”.

Page 2, line 21. Suggest “…raised the awareness of the problem in the remote sensing…”.

Page 2, line 22. Suggest “…follows their approach and introduces…”.

Page 2, line 24. Suggest a different word that methods as it was used at the beginning of the sentence.

Page 2, line 25. Suggest “…in each variable must be taken into account using approaches called errors-in-variables (EIV) regression.”

Page 2, line 30. Suggest remove “described.

Page 2, line 31. Suggest “ORDPACK is a somewhat…”.

Page 2, line 32. “Mahalanobis distance” is not a term most are familiar with. Might be worth a sentence and/or a reference to explain why it is different. Alternatively, perhaps leave out that detail.

Page 3, Lines 4-25. In discussing new particle formation rates and the relationship to sulphuric acid concentrations, the authors might consider discussion the following subjects:

Are the errors in measurement of J and H_2SO_4 related?

What is known about other factors that might affect the relationship between J and H_2SO_4 (such as water vapor, temperature, pressure, etc.)?

Page 3, Lines 4-11. See earlier comments about errors.

Page 3, line 12. Suggest “…particle formation rates at 1.7…”.

Page 3, line 13. Suggest “…concentrations simulated…”.

Page 3, line 13. Suggest “…pure sulphuric acid in nucleation experiments from the CLOUD…”.

Page 3, line 14. Suggest “…with corresponding expected values, their variances, and the covariance structures.”

Page 3, line 15-16. It is clear you are proud of the accomplishments using CLOUD, but this reviewer suggests removing the sentence that begins “The chamber data at CERN…”. Then, add CERN after “The” in the next sentence.

Page 3, line 18. The word precise is used twice in this sentence, but it does not say how precise. Given the earlier comments this reviewer made about the lack of direct connection between this study and NPF studies, perhaps the details of CERN and NPF studies could be reduced or eliminated (lines 15-20). In this discussion, the connection between $J_{1.7}$ and H_2SO_4 concentration is not clearly demonstrated. Is it not true that the calculation involves corrections for condensation and (for some sizes) wall loss? Suggest being more complete or leaving out this part.

Page 3, line 19. If this sentence remains in the paper, need another word or more discussion of what is meant by “inference”.

Page 3, line 13. Change : to ‘ after β.

Page 4, line 12. Suggest “In measured data, the variables…”

Page 4, line 13. Suggest “…the measurements, and the true….” and “Thus, we use simulated data…”

Page 4, line 15. Suggest “…formation rates ($J_{1.7}$) and sulphuric acid concentrations…”.

Page 4, line 20-21 and line 26. Suggest adding units to (molecules-cm^{-3}) to numbers.

Page 4, line 30. Suggest “This represents the quality…”.

Page 4. Before starting the Results section, suggest some discussion of the fit methods, perhaps in the supplement. Suggest adding some basic introduction to the fit methods in the paper. This reviewer suggests testing the application of all the methods by testing with a known data set, such as Pearson’s data with York’s weights (York, 1966) whose fit parameters are known with very high accuracy.

Page 5, line 8. It is not correct to say these methods had “equal accuracy” without stating the level of accuracy, in other words plus or minus an absolute level or plus or minus a percentage.

Page 5, line 11. From the errors given in Table 1, show how the totals errors used in Figure 2 were calculated.

Page 5, line 11. Suggest “…and with varying absolute and…”.

Page 5, line 14. Suggest “…significantly as more uncertainty…”.

Page 5, line 16. Suggest “…quite robust with increasing…”.

Page 5, line 17. Suggest “…of methods to decreasing number…”.

Page 5, line 20. Suggest “…estimated slopes can be very high.”

Page 5, line 20. Suggest "…slopes stabilize close to their characteristic levels (within xx% for five methods) for large datasets."

Page 5, line 21. Suggest "…more than 100 observations."

Page 5, line 22. It should be recognized that the number of points needed for a good fit depends on the uncertainties used. A few points will work fine if the uncertainties are small, while many more points are needed if uncertainties are large. This can perhaps be expressed at σ_x/x. Also, ensuring convergence is important for some of the methods (discussed above). To get an accurate representation of the data, it is also helpful for the data to cover a wide range. The x-data in this study only cover the range from about 5 to 7 ($\log_{10}[H_2SO_4]$). It would be interesting for fits when the values covered a factor of 5 to 10, even if they are not realistic for actual atmospheric situations.

Page 5, line 24. This reviewer was not sure what is meant by "high and low numbers" and "high number" in this sentence. This needs more discussion and clarity for the reader to understand clearly what was done.

Page 5, line 30. Suggest "…were not affected in either case…".

Page 5, line 31. Suggest "We did not explore how large a number of outliers would be needed to seriously disrupt the fits for the various methods. We felt that it is likely not realistic to have situations with more than 10% outliers.

Page 6, lines 2-4. This sentence needs rewording including improvement of the English to make it clear.

Page 6, line 4. Suggest "…of method should be based on the properties…".

Page 6, lines 5-8. This should be reworked based on suggestions made above.

Page 6, line 11-12. It states that the fits are made with "real" data. This is not true. These are all synthetic data. It also says that four of the methods gave slopes close to the true value. Suggest a quantitative comparison: slopes are within 5% of the true value (or whatever is appropriate). The methods are listed as good here are different than those listed in the Results section. Suggest making this consistent.

Page 6, line 14. It states that fits with small observations with all methods are highly uncertain. This does not agree with the earlier discussion and what is shown in Figure 3. Again, suggest quantitative comparisons and then statements about agreement (or lack of) that are also quantitative in this sentence and next few.

Page 6, line 15. Suggest "BLS was the most accurate…".

Page 6, line 16. Statement does not agree with the that made in Results.

Page 6, line 18. Suggest "…number of outliers (Figure 4), ODR and the York bivariate methods were the most stable…"

Page 6, line 20. Suggest "…sensitive to outliers after OLS."

Page 6, line 22. The recommendations depend on the level of uncertainty. Suggest being more quantitative, in other words, something like "When errors (σ_x/x) are greater than 50%, then method x and y performed systematically better than methods w and z."

Page 6, line 24. Suggest rewording "…we recommend considering twice…".

Page 6, line 25. Suggest "…robust with small numbers of data points." (Is this is what is meant?)

Page 6, line 32. Suggest "…were responsible for investigation…".