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.

A: We thank the reviewer for the helpful comments and suggestions. Our answers to the concerns addressed are below. The questions/comments from the reviewer are in italic font and our answers follow with plain text.

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.

A: In this response we will elaborate more the reasons why all suggested changes were not made. The English is also checked more carefully.

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.

A: In our opinion, technical notes are supposed to be short, concise pieces of information, which can be then applied for use in wider studies. We have added important information to text and supplement, with the help of comments of the reviewer, but we do not want to lengthen the manuscript too much.

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 x 10^6 molecules-cm^{-3}. The standard deviation is 2.4 x 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 x 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.

A: The selected simulation method resulted in distributions that are observed in atmosphere, e.g. for H_2SO_4 the distributions were similar as in our previous study (Mikkonen et al. ACP, 2011, 11, 11319-11334. doi:10.5194/acp-11-11319-2011). When simulating the data, negative values were immediately replaced by new simulated values until only non-negative values existed, which indeed offsets the simulated data from perfectly symmetric distributions in Fig 1. However, perfect symmetry is typically not a requirement for real datasets, hence we believe it should not be the case for our simulated data set either. It is obvious, that the shape of the data distributions will affect the performance of various regression estimators, but studying all possible related nuances would be a very laborious task. After all, the goal of this paper is to provide a simple demonstration of a frequently disregarded statistical phenomenon, not to exhaust all its possible variations that readers may encounter.

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.

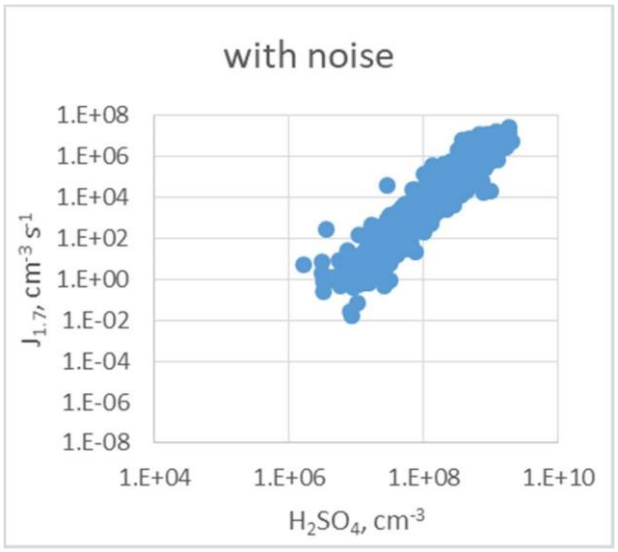

Figure R1. Synthetic data generated evenly spaced in logarithm space. Includes noise added to H_2SO_4 and $J_{1.7}$ values as described in Mikkonen et al., revised. OLS fit yields a slope of about 3.1, an intercept of about -21, and r^2 of about 0.9.

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).

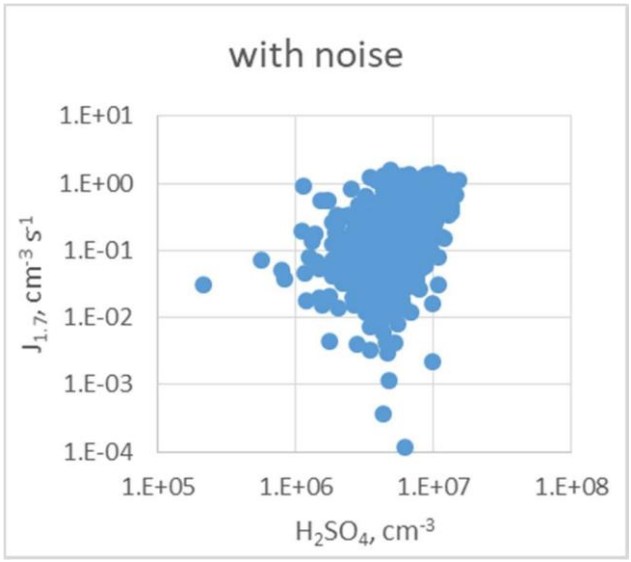

Figure R2. Synthetic data generated as in Figure R1, but with minimum and maximum ln(H_2SO_4) =15 and 16, respectively. OLS fit yields slopes of about 1.2, intercepts of about -9, and r^2 of 0.2. This appears to be caused by data covering a narrow dynamic range compared to the noise level.

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.
A: We thank the reviewer about the suggestion for new data simulation method. The method would assume that H_2SO_4 measurement would produce uniformly distributed data, which is not the case. Due to this, and other reasons listed above, we will keep the simulation method as is.

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.
A: In fact, Fig. 3 already visualizes the information on the variability of the slope estimate with repeated random datasets. For instance, the rightmost points on the figure show the medians and the shading shows the 1 std range of slope values for 40 random subsamples with 500 values in each sample. At 500 values per sample, the variability of the slope is already quite small and doubling the sample size to 1000, as the reviewer suggests, would further narrow the variability. In our view, Fig. 3 already gives a good visual understanding of the range of slope 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.
A: It is true that we have access to these type of data but currently, as the data are not open, we do not have permission to use them in this study. We have made tests with data collected from Hyytiälä and San Pietro Capofiume, Italy and the results are similar than seen here with simulated data. However, we cannot publish these results and thus we will publish a Python tool for running some of the methods and encourage people to test that on their own data. The tool can be found in GitHub: https://gist.github.com/mikkopitkanen/da8c949571225e9c7093665c9803726e
The link is also added to the end of the manuscript.

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.
A: We pointed out the effect of weighting in Conclusions section and give a reference to study demonstrating effects of it with sentence: “Weighting the data based on some factor, typically the inverse

of the uncertainty, reduces the effect of outliers and makes the regression depend more on the data that is more certain (see e.g. Wu and Yu, 2018) but it does not solve the problem completely."

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.

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.

A: We agree with the reviewer that the selection of the convergence criteria is important in order to get reliable results. However, we feel that for the ready-made packages e.g. in R we can trust the developers that they have set the criteria on optimal level. In addition, we feel that every researcher conducting these kind of analyses should by default be aware of the nature of iterative methods and issues with convergence. These are taught in basic courses of data analysis. We will add description of York method in the supplement and clarify other descriptions of methods where needed. The tool for applying the methods will also be available, as stated in comment above. Also, substantial number of tests with known data were conducted with the custom York (2004) implementation to ensure that the program works as expected by using simulated data but also the so called Pearson's data set shows for example in Cantrell (2008) Fig 1. The source code already published on Github allows readers to verify the code themselves:
https://gist.github.com/mikkopitkanen/da8c949571225e9c7093665c9803726e

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.

A: We thank the reviewer for pointing out the typo in the equation, which is now corrected in the revised supplement.

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.

A: In the simulation of the data, the values were generated one by one and each negative value was immediately replaced with a new simulated value, until only non-negative values were included in the data set. This is why no negative data values exist in the first place and the conversion to logarithmic scale did not eliminate any data. Still, the referee is correct, that even with this method significant number of negative values would cause bias but in our data the number was so small that the effect of negligible

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".

A: Term "homoscedasticity" replaced with "equal variances". Terms "heteroscedasticity" and "a posteriori" defined in section 2.2. Terms "estimators" and "predictors" should be known by all who are conducting regression analysis.

Manuscript specific comments

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

A: Corrected as suggested

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…"

A: This would change the meaning of the sentence. It is supposed to indicate that line fitting is not that simple as people usually think.

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

A: Corrected as suggested

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

A: Sentence changed to form: "If the relationship is not defined correctly, the inference is not valid either."

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

A: The sentence changed to form: “Regression models can be linear or non-linear, depending on the relationship between data sets that are analysed.”

Page 1, lines 26-27. Suggest “…that the independent variable of the model has been measured without error and the model accounts only for…”
A: Corrected as suggested

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.
A: The asymptotics here refer to number of data. For clarity, the word “asymptotically” is replaced as: “with very high number of number of data points”

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.
A: The sentences reformulated as: “If predictor variables in regression analysis contain any measurement error, methods that account for errors should be applied. Particularly when errors are large. Thus, test variables in this study were chosen such that they 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…”
A: Corrected as suggested

Page 2, line 11. Suggest adding information to describe what is meant by “unconstrained” in this context.
A: We are not certain which unconstrained method was used, a note on this was added to text.

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…”
A: Corrected as suggested

Page 2, line 26. EIV methods simply mean that errors in both variables are accounted for. Suggest a statement that says this.
A: Added sentence: “EIV methods simply mean that errors in both variables are accounted for.”

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.
A: The sentence is comparing ORDPACK only to classical orthogonal regression. For clarity, the sentence is reworded as: “ORDPACK is a somewhat improved version of classical orthogonal regression,

so that arbitrary covariance structures are acceptable and is specifically set up so that a user can specify measurement error variances and covariance point by point, as some of the methods in this study are doing in linear analysis."

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.

A: In the actual line fitting, the bias cannot be taken account but it needs to be minimized, as the reviewer noted, with calibrations and zeros or by data pre-processing. For clarity, the sentence is reworded as: "In line fitting, bias cannot be taken account but it needs to be minimized through careful and regular instrument calibrations and zeros or data pre-processing."

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 gasphase 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…"

A: Corrected as suggested

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.

A: The sentence is changed to form "However, using ODR allows for performing regression on a user defined model, while the York (2004) solution works only on linear models. This, for instance, enables using linear scale uncertainties in ODR in this study, while the York (2004) approach could only use log scale uncertainties."

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.

A: Corrected as suggested

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.

A: The sentence reworded as "PCA accounts only for the observed variance in data, whereas ODR, Bayes EIV and York bivariate regression require known estimates for measurement errors."

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

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

A: Corrected as suggested

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.

A: Corrected as suggested

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.

A: Corrected to lower case

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.

A: The sentence removed as suggested

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.

A: "true" changed as "noise-free" as suggested in later comment

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

Page 5, line 32. Outliers are also generated on the low tail of the distribution. These should be eliminated as well.
A: This is how it was done; we clarified it to text by talking about absolute error

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."
A: Corrected as suggested

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.
A: The sentence is changed to form: "It shows that when the uncertainty is small, the bias in OLS fit is smaller but when more uncertainty is added to data the bias increases significantly"

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…"
A: Corrected as suggested

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.
A: We elaborated the recommendation to cases with similar uncertainty by adding this to the end of the sentence: "…and similar type of uncertainty than presented here"

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

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.
A: The sentence reworded as: "…because the methods behave slightly differently with different types of error"

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.
A: The sentence is removed from the manuscript.

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.

A: The sentence refers to simulated data in Fig 3. It is true that for single dataset methods can be exact but even when the sample in the analysis is drawn from the same distribution containing some error, the methods will have uncertainty in the analysis. Some methods are more robust for this than others. Already Cantrell (2008) stated that for methods recommended there: "the accuracy of the slope improves with the number of data points (not so with the standard least squares with significant errors in the x-variable)" The sentence is reformulated as "The main message to learn in Fig 3 is that if the data contain some error, then with small numbers of observations all fit methods are highly uncertain."

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.

A: symbol defined as "relative error" already in the manuscript line 29. This is the first time this measure is needed.

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.

A: The sentence changed to form "with chosen uncertainties in our simulation tests BLS showed out to be the most robust with small numbers of data points."

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.

A: We reworded both the sentence in the conclusions and the supplement section regarding the Bayes EIV method and provided with additional symbol definitions.

Supplement specific comments.

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

A: The variables are now defined and the equations numbered.

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.

A: In the beginning, we added sentence "In this supplement, we introduce the minimizing criteria (C_{method}) for all methods applied in the main text. We also give the equations for regression coefficients ($\hat{\alpha}_{method}$ and $\hat{\beta}_{method}$) for the methods." In addition, headers are added for each method.

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?

A: The sentence corrected to form: "ODR takes into account that errors exist in both axes but not the exact values of the variances of variables."

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.

A: Corrected as suggested

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.

A: The equation is corrected as stated above.

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

A:Corrected as suggested

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

A: The difference is in the coefficients $\hat{\alpha}$ and $\hat{\beta}$, which are defined below.

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

A:Corrected as suggested

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.)

A: Corrected as suggested. The convergence criteria are given in Stan documentation and, as stated above, we trust the software developers that the criteria are valid.

Technical note: Effects of Uncertainties and Number of Data points on Line Fitting - a Case Study on New Particle Formation

Santtu Mikkonen[1], Mikko R. A. Pitkänen[1,2], Tuomo Nieminen[1§], Antti Lipponen[2], Sini Isokääntä[1], Antti Arola[2], and Kari E. J. Lehtinen[1,2]

[1] Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[2] Finnish Meteorological Institute, Atmospheric Research Centre of Eastern Finland, Kuopio, Finland
[§] Currently at: Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland

Correspondence to: Santtu Mikkonen (santtu.mikkonen@uef.fi)

Abstract. Fitting a line ~~on a scatterplot~~ of two measured variables is considered ~~as~~ one of the simplest statistical procedures researchers can do. However, this simplicity is deceptive as the line fitting procedure is actually quite a complex problem. Atmospheric measurement data never comes without some measurement error. Too often, these errors are neglected when researchers are making inferences from their data.

To demonstrate the problem, we simulated datasets with different amounts of data and error, mimicking the dependence of atmospheric new particle formation rate ($J_{1.7}$) on sulphuric acid concentration (H_2SO_4). Both variables have substantial measurement error and thus they are good test variables for our study. We show that ordinary least squares (OLS) regression results in strongly biased slope values compared with six error-in-variables (EIV) regression methods (Deming, Principal component analysis, orthogonal, Bayesian EIV, and two different bivariate regression methods) known to take into account errors in the variables.

1 Introduction

Atmospheric measurements ~~data never~~ always come with~~out~~ some measurement error. Too often, these errors are neglected when researchers are making inferences based on their data. Describing the relationship between two variables typically involves making deductions in some more general context than was directly studied. If the relationship is not defined correctly ~~ill formulated~~, the inference is not valid either. In some cases, the bias in analysis method is even given a physical meaning. When analysing dependencies of two or more measured variables, regression models are usually applied. Regression models can be linear or non-linear, depending on the relationship between data sets that are analysed. ~~A regression model can be linear or nonlinear, depending on the data~~. Standard regression models assume that the independent variables of the model~~s~~ have been measured without error and the model~~s~~ account only for errors in the dependent variables or responses. In cases where the measurements of the predictors contain error, estimating with standard methods, usually Ordinary Least Squares (OLS), do not tend to the true parameter values, not even with very high number of number of data points ~~asymptotically~~. In linear

models, the coefficients are underestimated (e.g. Carroll et al., 2006) but in nonlinear models, the bias is likely to be more complicated (e.g. Schennach 2004). If predictor variables in regression analysis contain any ~~M~~measurement error, methods that account for errors should be applied. ~~needs to be taken into account, particularly~~ Particularly when errors are large. Thus, ~~we chose such parameters as our~~ test variables in this study were chosen such that they included significant uncertainties in both the independent and dependent variables. Sulphuric acid (H_2SO_4) is known to strongly affect the formation rates (J) of aerosol particles (Kirkby et al., 2016; Kuang et al., 2008; Kulmala et al., 2006; Kürten et al., 2016; Metzger et al., 2010; Riccobono et al., 2014; Riipinen et al., 2007; Sihto et al., 2006; Spracklen et al., 2006). The relationship between J (cm^{-3} s^{-1}) and H_2SO_4 (molec cm^{-3}) is typically ~~as~~ assumed to be in form $\log_{10}(J) = \beta * \log_{10}(H_2SO_4) + \alpha$ (Seinfeld and Pandis, 2016). In addition, parameterizations based on the results from these fits have been implemented in global models, e.g. in (Dunne et al., 2016; Metzger et al., 2010; Spracklen et al., 2006), to estimate the effects of new particle 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 first nucleation theorem (Vehkamäki, 2006)..

Some published results have shown discrepancies in the expected J vs H_2SO_4 dependence. Analysing data from Hyytiälä in 2003, Kuang et al. (2008) used an unconstrained least squares method, which was not specified in the paper, and obtained β=1.99 for the slope, whereas Sihto et al. (2006) reported a value of β=1.16 using OLS from the same field campaign. They had some differences in pre-treatment of data and used different time windows, but a significant proportion of this inconsistency is very likely due to use of different fitting methods~~is very likely due to different methods for making the fit~~. The problem in the relationship of H_2SO_4 and J has been acknowledged previously ~~already~~ in Paasonen et al. (2010) who noted that bivariate fitting method as presented in York et al. (2004) should be applied but could not be used due to the lack of proper error estimates for each quantity. They were not aware of the methods that do not need to know the errors in advance, but instead made use of estimated variances. Here, we present appropriate tools for using that approach.

Multiple attempts have been made to present methods accounting for errors in predictor variables for regression-type analysis, going back to Deming (1943). However, the traditional least squares fitting still holds the position as the de facto line fitting method due to its simplicity and common availability in frequently used software. In atmospheric sciences, Cantrell (2008) drew attention to the method introduced by York (1966) and York et al. (2004) and listed multiple other methodological papers utilizing similar methodology. Pitkänen et al. (2016) raised the awareness of the problem in remote sensing community and this study partly follows their approach and introduces multiple methods to take account the errors in predictors. Cheng and Riu (2006) studied methods with heteroscedastic errors whereas Wu and Yu (2018) approached the problem with measurement errors via weighted regression and applied some techniques also used in our study.

Measurement errors in each variable must be taken into account using approaches called errors-in-variables (EIV) regression. EIV methods simply mean that errors in both variables are accounted for. In this study, we compared OLS regression results to six different regression methods (Deming regression, Principal component analysis regression, orthogonal regression, Bayesian EIV regression and two different bivariate regression methods) known to be able to take into account errors in variables and provide (at least asymptotically) unbiased estimates. In this study, we will focus only on linear EIV methods but

it is important to acknowledge that there also exist nonlinear methods e.g. ORDPACK introduced in Boggs, Byrd, and Schnabel (1987) and implemented in Python SciPy and R (Boggs et al., 1989; Spiess, 2015). ORDPACK is a somewhat improved version of classical orthogonal regression, so that arbitrary covariance structures are acceptable and is specifically set up so that a user can specify measurement error variances and covariance point by point, as some of the methods in this study are doing in linear analysis.

2 Materials and Methods

2.1 Data illustrating the phenomenon

Measurement data contains different types of errors. Usually, the errors are divided to two main class: random and systematic error. Systematic errors, commonly referred as bias, in experimental observations usually come from the measuring instruments. They may occur because there is something wrong with the instrument or its data handling system, or because the instrument is not used correctly by the operator. In line fitting, bias cannot be taken account but it needs to be minimized through careful and regular instrument calibrations and zeros or data pre-processing. ~~the~~ The random error instead may have different components, of which two are discussed here: natural error and measurement error. In addition, one should note the existence of equation error, discussed in Carroll and Ruppert (1996), which refers to using an inappropriate form of a fitting equation. Measurement error is more generally understood, it is where measured values do not fully represent the true values of the variable being measured. This also contains sampling error, e.g. in the case of H_2SO_4 measurement the sampled air in the measurement instrument is not representative sample of outside air (e.g. due to losses of H_2SO_4 occurring in the sampling lines). Natural error is the variability caused by natural or physical phenomenon~~Natural error is that the true connection between the two variables is has stochastic variation by some natural or physical cause~~ e.g. certain amount of H_2SO_4 does not cause same number of new particles formed. In the analysis of measurement data, some amount of these errors are known or can be estimated, but some of it will usually remain unknown, which should be kept in mind when interpreting ~~data~~fits. Even though the measurement error is taken into account, the regression fit may be biased due to unknown natural error. In this study, we assume that the errors of the different variables are uncorrelated, but in some cases ~~it~~ this has to be taken into account, as noted e.g. in Trefall and Nordö (1959) and Mandel (1984). The correlation between the errors of two variables, measured with separate instruments, independent ~~on~~ of each other, like formation rate and H_2SO_4, may come e.g. from environmental variables affecting both of them at the same time. Factors affecting formation of sulphuric acid have been studied in various papers, e.g. in Weber et al. (1997) and Mikkonen et al. (2011). New particle formation rates, in turn, have been studied e.g. in Boy et al.(2008) and in Hamed et al. (2011) and similarities between affecting factors can be seen. In addition, factors like room temperature in the measurement space and atmospheric pressure may affect the performance of instrumentation~~room~~

~~temperature in measurement space and atmospheric pressure may affect to measurement instruments~~, thus causing additional error.

The data used in this study consist of simulated new particle formation rates at 1.7 nanometre size ($J_{1.7}$) and sulphuric acid (H_2SO_4) concentrations mimicking observations of pure sulphuric acid in nucleation experiments from the CLOUD chamber in CERN (Kürten et al. 2016; https://home.cern/about/experiments/cloud) with corresponding expected values, their variances and covariance structures. ~~The chamber data at CERN are the best characterized and controlled set of new particle formation (NPF) experiments in the history of aerosol science so far.~~ The Proton Synchrotron provides an artificial source of "cosmic rays" that simulates natural conditions of ionization between ground level and the stratosphere. The core is a large (volume 26m3) electro-polished stainless steel chamber with temperature control (temperature stability better than 0.1 K) at any tropospheric temperature, precise delivery of selected gases (SO_2, O_3, NH_3, various organic compounds) and ultrapure humidified synthetic air, and very low gas-phase contaminant levels. The existing data on NPF includes what are believed to be the most important routes that involve sulphuric acid, ammonia and water vapour~~Existing data include the most suspected candidates for atmospheric NPF, including sulphuric acid – ammonia – water~~ (Kirkby et al., 2011), sulphuric acid – amine (Almeida et al., 2013) and ion induced organic nucleation (Kirkby et al., 2016).The actual nucleation of new particles occurs at slightly smaller size. After formation, they grow by condensation to reach the detection limit (1.7 nm) of the instrument and $J_{1.7}$ thus refers to the formation rate of particles as the instrument detects them, taking into account the known particle losses due to coagulation and deposition on the chamber walls. The relationships between precursor gas phase concentrations and particle formation rates ~~These variables~~ were chosen because they are both known to have considerable measurement errors and their relationship is studied frequently using regression-based analyses (Kirkby et al., 2016; Kürten et al., 2016; Riccobono et al., 2014; Tröstl et al., 2016)~~ which makes them good illustrative variables for this study~~. Additionally, many of the published papers on this topic do not describe how they are taking account the uncertainties in the analysis, which casts doubt that errors have been treated properly~~leaves a doubt that they are not treated properly~~. However, it should be kept in mind that the data could be any set of numbers assumed to have 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. ~~but to raise the awareness in the research community we related the simulations to well-known datatype.~~

2.2 Regression methods

We made fits for the linear dependency of logarithms of the two study variables, such that the equation for the fit was given by

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1)$$

where y represents $\log_{10}(J_{1.7})$, x is $\log_{10}(H_2SO_4)$, β's are the coefficients estimated from the data and ε is the error term. In order to demonstrate the importance of taking into account the measurement errors in the regression analysis, we tested seven different line-fitting methods. Ordinary Least Squares (OLS), not taking account the uncertainty in x-variable, and orthogonal regression (ODR, Boggs, Byrd, and Schnabel 1987), Deming regression (DR, Deming, 1943), Principal component analysis

(PCA, Hotelling, 1957) regression, Bayesian EIV regression (Kaipio and Somersalo, 2005) and two different bivariate least squares methods by York *et al.*, (2004), and Francq and Govaerts (BLS, 2014), known to be able to take account errors in variables and provide (at least asymptotically) unbiased estimates. The differences between the methods come from the criterion they minimize when calculating the coefficients and how they ~~take~~ account for ~~the~~ measurement errors. The minimizing criteria for all methods are given in ~~the~~ supplement S1, but here we give ~~here~~ the principles of the methods. OLS minimizes the squared distance of the observation and the fit line either in y or x direction, but not both at the same time, whereas ODR minimizes the sum of squared weighted orthogonal distances between each point and the line. DR was originally an improved version of orthogonal regression, taking account the ratio of the error variances, λ_{xy}, of the variables, (in classical non-weighted ODR λ_{xy} =1) and it is the maximum likelihood estimate (MLE) for the model (1) when λ_{xy} is known. The approach ~~Idea~~ of PCA is the same as in ODR but the estimation procedure is somewhat different as can be seen in S1. The bivariate algorithm by York et al 2004 provides a simple set of equations for iterating MLE of slope and intercept with weighted variables, which makes it similar to ODR in this case. However, using ODR ~~can consider linear scale uncertainties in logarithmic scale regression, unlike~~allows for performing regression on a user defined model, while the York (2004) solution works only on linear models. This, for instance, enables using linear scale uncertainties in ODR in this study, while the York (2004) approach could only use log scale uncertainties. In Bayes EIV, statistical models for the uncertainties in observed quantities are used and probability distributions for the line slope and intercept are computed according to the Bayes' theorem. In this study, we computed the Bayesian maximum a posteriori (MAP) estimates for the slope and intercept that are the most probable values given the likelihood and prior models, see Supplement S1 for more details on models used in Bayes EIV. ~~In this study, we computed Bayesian maximum a posteriori (MAP) estimates for the slope and intercept values.~~ BLS takes into account errors and heteroscedasticity, i.e. unequal variances, in both ~~axes~~ variables and thus is more advanced method than DR (under normality and ~~homoscedasticity~~equal variances, BLS is exactly equivalent to DR). PCA accounts only for the ~~measurement~~ observed variance in data, whereas ODR, Bayes EIV and York bivariate regression require known estimates for measurement errors. Though~~t~~ for Bayes EIV the error can be approximated with a distribution. DR and BLS can be applied with both~~,~~ errors given by the user and measurement variance based errors. In this study, we applied measurement variance based errors for them. The analysis for OLS and PCA were calculated with R-functions “lm” and “prcomp”, respectively (R Core Team, 2018) DR was calculated with package “deming” (Therneau, 2018) and BLS with package “BivRegBLS” (Francq and Berger, 2017) in R. The ODR based estimates were obtained using “scipy.odr” python package (Jones et al., 2001), while the python package “pystan” (Stan Development Team, 2018) was used for calculating the Bayesian regression estimates. Finally, the York bivariate estimates were produced with a custom python implementation of the algorithm presented by York et al. (2004).

3 Data simulation

In measured data, the variables that are observed are not x and y, but $(x+e_x)$ and $(y+e_y)$, where e_x and e_y are the uncertainty in the measurements, and the true x and y cannot be exactly known. Thus, we used simulated data, where we know the true, i.e. noise-free x and y, to illustrate how the different line fitting methods perform in different situations.

We simulated a dataset mimicking new particle formation rates ($J_{1.7}$) and sulphuric acid concentrations (H_2SO_4) reported from CLOUD-chamber measurements in CERN. Both variables are known to have substantial measurement error and thus they are good test variables for our study. Additionally, the relationship of logarithms of these variables is quite often described with linear OLS regression and thus the inference may be flawed.

We generated one thousand random noise-free ~~"true"~~ H_2SO_4 concentration values assuming log-normal distribution with median $2.0*10^6$ *(molecules cm^{-3})* and standard deviation $2.4*10^6$ *(molecules cm^{-3})*. The corresponding noise-free ~~true~~ $J_{1.7}$ was calculated using model $\log_{10}(J_{1.7}) = \beta*\log_{10}(H_2SO_4)+\alpha$ with the noise-free ~~true~~ slope β=3.3 and α=*-23*, both are realistic values presented by Kürten *et al.* (2016, Table 2 for the no added ammonia cases). ~~The resulting $J_{1.7}$ mean was 0.11 and standard deviation was 0.50, similar to $J_{1.7}$ statistics in Kürten et al. (2016).~~

Simulated observations of the noise-free ~~true~~ H_2SO_4 were obtained by adding random errors $e_x = e_{rel,x}x + \sigma_{abs,x}$ that have a random absolute component $e_{abs,x} \sim \text{normal}(0,\sigma_{abs,x})$ and a random component relative to the observation x itself $e_{rel,x}x$, where $e_{rel,x} \sim \text{normal}(0,\sigma_{rel,x})$. Similar definitions apply for the noise-free ~~true~~ $J_{1.7}$, e_y, $\sigma_{abs,y}$ and ~~σ_{abs}~~$\sigma_{rel,y}$. The standard deviations of the measurement error components were chosen $\sigma_{abs,x} = 4*10^5$, $\sigma_{rel,x} = 0.3$, $\sigma_{abs,y} = 3*10^{-3}$, $\sigma_{rel,y} = 0.5$, which are subjective estimates based on measurement data.

Simulating the observations tends to generate infrequent extreme outlier observations from the infinite tails of the normal distribution. We discarded these outliers with ~~an~~ absolute error larger than three times the combined standard uncertainty of the observation in order to remove the effect of outliers from the regression analysis. This represents the quality control procedure in data analysis and it also improved the stability of our results between different simulations.

4 Results

Differences between the regression methods are illustrated with four different ways. Firstly, by showing line fits on scatterplot of simulated data. Secondly, illustrating how the slopes change when the uncertainty in the measured variables increase, thirdly by showing the sensitivity of the fits on number of observations and finally showing how the fits are affected by adding outliers in the data.

Regression fits with all methods in use are shown in Figure 1. As we know that the ~~"true slope"~~ noise-free slope β_{true}=3.30 we can easily see how the methods perform. The worst performing method was OLS, with β_{ols}=1.55, which is roughly half of the β_{true}. The best performing methods with equal accuracy, i.e. within 2% range, were ODR (β_{ODR}=3.27), Bayes EIV (β_{BEIV}=3.24) and BLS (β_{BLS}=3.22), whereas York (β_{York}=3.15) was within 5% range, but Deming (β_{DR}=2.95) and PCA (β_{PCA}=2.92) slightly underestimated the slope.

The sensitivity of the methods was first tested by varying the uncertainty in H_2SO_4 observations. We simulated six datasets with 1000 observations and with varying absolute and relative uncertainties, listed in Table 1, and performed fits with each method on all of these datasets~~made fits with each method on all datasets separately~~. The performance of the methods is shown in Figure 2, with the results corresponding to Figure 1 ~~are~~ marked ~~with~~ in black ~~colour~~. It shows that when the uncertainty is small~~er~~, the bias in OLS fit is smaller but when more uncertainty is added to data the bias increases significantly ~~as more uncertainty is added to data~~. Decrease in performance can also be seen with ODR, which ~~is overestimating~~ overestimates the slope, and PCA, DR and Bayes EIV, which all underestimate the slope. Bivariate methods, BLS and York, seem to be quite robust with increasing uncertainty, as the slopes are not changing ~~considerably~~significantly.

The sensitivity of methods to decreasing number of observations was tested by picking 100 random samples from the 1000 simulation dataset with n of 3, 5, 10, 20, 30, 50, 70, 100, 300 and 500 and making fits for all samples with all methods. The average slopes and their standard errors are shown in Figure 3. It is clear that when the number of observations is 10 or less, the variation in estimated slopes can be considerably high. When $n \geq 30$ the average slopes stabilized close to their characteristic levels (within 5%), except for Bayes EIV and York bivariate, which needed more than 100 observations. The most sensitive methods for small n ~~are~~ were Bayes EIV, ODR and PCA and thus they should not be applied for data with small n and similar type of uncertainty than presented here. Though, it should be remembered that number of points needed for a good fit depends on the uncertainties in the data.

The sensitivity for outliers in predictor variable H_2SO_4 was tested with two different scenarios. First, the outliers were let to be randomly either high or low end of the distribution. In the second scenario, outliers were allowed to be only large numbers, which is often the case in H_2SO_4 and aerosol concentration measurements as the smallest numbers are cleaned out from the data when they are smaller than the detection limit of the measurement instrument. Five cases with n=1000 were simulated with increasing number of outliers (0, 5, 10, 20, 100) and 10 repetitions of H_2SO_4 values with different set of outliers. Outliers were defined such that x_{obs}-x_{true}>3*combined standard uncertainty. The most sensitive methods for outliers in both scenarios were OLS and Bayes EIV. High number of outliers caused underestimation to PCA and DR, especially in high outlier case, and slight overestimation to BLS in random outlier case. York Bivariate and ODR were not affected in either case and BLS had only small variation between the 10 replicates in the estimated slope. 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.

5 Conclusions

Ordinary least squares regression can be used to answer some simple questions on data, such as "How is Y related to X?". However, if we are interested in the strength of the relationship and the predictor variable X contains some error, then error-in-variables methods should be applied. There is no single correct method to make the fit, because the methods behave slightly differently with different types of error ~~measure slightly different things about the data~~. The choice of method should be based

on the properties of data and the specific research question. There are usually two types of error in the data: natural and measurement error, where natural error refers to stochastic variation in the environment. Even if the natural error in the data is not known, taking into account the measurement error improves the fit significantly. Weighting the data based on some factor, typically the inverse of the uncertainty, reduces the effect of outliers and makes the regression depend more on the data that is more certain (see e.g. Wu and Yu, 2018) but it does not solve the problem completely. ~~In addition, no matter how small the measurement error would be, it should be taken account because taking it into account will never lead to more biased estimator.~~

As a case study, we simulated a dataset mimicking the dependence of atmospheric new particle formation rate on sulphuric acid concentration. We introduced three major sources of uncertainty when doing inference from scatterplot data: increasing measurement error, number of data points and number of outliers. In Fig 1, we showed that in case of simulations where errors are taken from real measurements of $J_{1.7}$ and H_2SO_4 four of the methods gave slopes within 5% of the ~~"true"~~ known noise-free value: BLS, York bivariate, Bayes EIV and ODR. Estimates from BLS and York bivariate remained stable even when the uncertainty in simulated H_2SO_4 was increased drastically in Fig 2. The main message to learn in Fig 3 is that if the data contain some error, then with small numbers of observations all fit methods are highly uncertain. BLS was the most accurate with smallest sample sizes of 10 and less, ODR stabilized with 20 observations and York bivariate and Bayes EIV needed 100 or more data points to become accurate. After that, they approach the noise-free ~~true~~ value asymptotically, while the OLS slope, in contrast, converges towards an incorrect value. With the increasing number of outliers (Figure 4) ODR and York bivariate were the most stable ones, even with 10% of observations classified as outliers in both test cases. BLS remained stable in the case with only high outliers. Bayes EIV was the most sensitive to outliers after OLS.

From this, we can give a recommendation that if the uncertainty in predictor is known, York bivariate, or other method able to use known variances, should be applied. If the errors are not known, and they are estimated from data, BLS and ODR showed out to be the most robust in cases of increasing uncertainty (relative error rE > 30% in Fig 2) and with high number of outliers. In our test data, BLS and ODR stayed stable up to rE >80% in Fig. 2 whereas DR and PCA started to be more uncertain when rE > 30% and Bayes EIV when rE>50%. If the number of observations is less than 10, and the uncertainties are high, we recommend considering if a regression fit is appropriate at all. However, with chosen uncertainties in our simulation tests BLS showed out to be the most robust with small numbers of data points. Bayes EIV has significant advantages if the number of observations is high enough and there are not too many outliers, as it ~~is able to estimate the errors in data with distributions.~~ does not require explicit definition of the errors but can treat them as unknown parameters given their probability distributions.

Author contribution

SM prepared the manuscript with contributions from all co-authors. SM, MP and SI performed the formal analysis. MP simulated the data. SM, AA and KL formulated the original idea. SM, MP and AL developed and implemented the methodology. SM, MP, TN and AL were responsible for investigation and validation of data and methods.

Acknowledgments

This work was supported by The Nessling foundation and The Academy of Finland Centre of Excellence (grant no. 307331).

Competing interests

The authors declare that they have no conflict of interest.

Code availability: Python code for running the methods can be found in GitHub: https://gist.github.com/mikkopitkanen/da8c949571225e9c7093665c9803726e

Data availability: Simulated datasets used in the example analysis will be given as supplement upon publication.

6 References

Almeida, J., Schobesberger, S., Kürten, A., Ortega, I. K., Kupiainen-Määttä, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurtén, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppä, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petäjä, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamäki, H. and Kirkby, J.: Molecular understanding of sulphuric acid–amine particle nucleation in the atmosphere, Nature, 502(7471), 359–363, doi:10.1038/nature12663, 2013.

Boggs, P. T., Byrd, R. H. and Schnabel, R. B.: A Stable and Efficient Algorithm for Nonlinear Orthogonal Distance Regression, SIAM J. Sci. Stat. Comput., 8(6), 1052–1078, doi:10.1137/0908085, 1987.

Boggs, P. T., Donaldson, J. R., Byrd, R. H. and Schnabel, R. B.: Algorithm 676 ODRPACK: software for weighted orthogonal distance regression, ACM Trans. Math. Softw., 15(4), 348–364, doi:10.1145/76909.76913, 1989.

Boy, M., Karl, T., Turnipseed, A., Mauldin, R. L., Kosciuch, E., Greenberg, J., Rathbone, J., Smith, J., Held, A., Barsanti, K., Wehner, B., Bauer, S., Wiedensohler, A., Bonn, B., Kulmala, M. and Guenther, A.: New particle formation in the Front Range of the Colorado Rocky Mountains, Atmos. Chem. Phys., 8(6), 1577–1590, doi:10.5194/acp-8-1577-2008, 2008.

Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8(17), 5477–5487, doi:10.5194/acp-8-5477-2008, 2008.

Carroll, R. J. and Ruppert, D.: The Use and Misuse of Orthogonal Regression in Linear Errors-in-Variables Models, Am. Stat., 50(1), 1–6, doi:10.1080/00031305.1996.10473533, 1996.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M.: Measurement error in nonlinear models : a modern

perspective., 2nd Editio., Chapman & Hall/CRC., 2006.

Cheng, C.-L. and Riu, J.: On Estimating Linear Relationships When Both Variables Are Subject to Heteroscedastic Measurement Errors, Technometrics, 48(4), 511–519, doi:10.1198/004017006000000237, 2006.

Deming, W. E.: Statistical adjustment of data, Wiley, New York., 1943.

Dunne, E. M., Gordon, H., Kürten, A., Almeida, J., Duplissy, J., Williamson, C., Ortega, I. K., Pringle, K. J., Adamov, A., Baltensperger, U., Barmet, P., Benduhn, F., Bianchi, F., Breitenlechner, M., Clarke, A., Curtius, J., Dommen, J., Donahue, N. M., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Jokinen, T., Kangasluoma, J., Kirkby, J., Kulmala, M., Kupc, A., Lawler, M. J., Lehtipalo, K., Makhmutov, V., Mann, G., Mathot, S., Merikanto, J., Miettinen, P., Nenes, A., Onnela, A., Rap, A., Reddington, C. L. S., Riccobono, F., Richards, N. A. D., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Simon, M., Sipilä, M., Smith, J. N., Stozkhov, Y., Tomé, A., Tröstl, J., Wagner, P. E., Wimmer, D., Winkler, P. M., Worsnop, D. R. and Carslaw, K. S.: Global atmospheric particle formation from CERN CLOUD measurements., Science, 354(6316), 1119–1124, doi:10.1126/science.aaf2649, 2016.

Francq, B. G. and Berger, M.: BivRegBLS: Tolerance Intervals and Errors-in-Variables Regressions in Method Comparison Studies. R package version 1.0.0., [online] Available from: https://cran.r-project.org/package=BivRegBLS, 2017.

Francq, B. G. and Govaerts, B. B.: Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives, Chemom. Intell. Lab. Syst., 134, 123–139, doi:10.1016/j.chemolab.2014.03.006, 2014.

Hamed, A., Korhonen, H., Sihto, S.-L., Joutsensaari, J., Järvinen, H., Petäjä, T., Arnold, F., Nieminen, T., Kulmala, M., Smith, J. N., Lehtinen, K. E. J. and Laaksonen, A.: The role of relative humidity in continental new particle formation, J. Geophys. Res., 116(D3), D03202, doi:10.1029/2010JD014186, 2011.

Hotelling, H.: The Relations of the Newer Multivariate Statistical Methods to Factor Analysis, Br. J. Stat. Psychol., 10(2), 69–79, doi:10.1111/j.2044-8317.1957.tb00179.x, 1957.

Jones, E., Oliphant, T. and Peterson, P.: SciPy: Open Source Scientific Tools for Python, [online] Available from: http://www.scipy.org/, 2001.

Kaipio, J. and Somersalo, E.: Statistical and Computational Inverse Problems, Springer-Verlag, New York., 2005.

Kirkby, J., Curtius, J., Almeida, J., Dunne, E., Duplissy, J., Ehrhart, S., Franchin, A., Gagné, S., Ickes, L., Kürten, A., Kupc, A., Metzger, A., Riccobono, F., Rondo, L., Schobesberger, S., Tsagkogeorgas, G., Wimmer, D., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Downard, A., Ehn, M., Flagan, R. C., Haider, S., Hansel, A., Hauser, D., Jud, W., Junninen, H., Kreissl, F., Kvashin, A., Laaksonen, A., Lehtipalo, K., Lima, J., Lovejoy, E. R., Makhmutov, V., Mathot, S., Mikkilä, J., Minginette, P., Mogo, S., Nieminen, T., Onnela, A., Pereira, P., Petäjä, T., Schnitzhofer, R., Seinfeld, J. H., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Vanhanen, J., Viisanen, Y., Vrtala, A., Wagner, P. E., Walther, H., Weingartner, E., Wex, H., Winkler, P. M., Carslaw, K. S., Worsnop, D. R., Baltensperger, U. and Kulmala, M.: Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation, Nature,

476(7361), 429–433, doi:10.1038/nature10343, 2011.

Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., Heinritzi, M., Simon, M., Yan, C., Almeida, J., Tröstl, J., Nieminen, T., Ortega, I. K., Wagner, R., Adamov, A., Amorim, A., Bernhammer, A.-K., Bianchi, F., Breitenlechner, M., Brilke, S., Chen, X., Craven, J., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Hakala, J., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Molteni, U., Onnela, A., Peräkylä, O., Piel, F., Petäjä, T., Praplan, A. P., Pringle, K., Rap, A., Richards, N. A. D., Riipinen, I., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Scott, C. E., Seinfeld, J. H., Sipilä, M., Steiner, G., Stozhkov, Y., Stratmann, F., Tomé, A., Virtanen, A., Vogel, A. L., Wagner, A. C., Wagner, P. E., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Zhang, X., Hansel, A., Dommen, J., Donahue, N. M., Worsnop, D. R., Baltensperger, U., Kulmala, M., Carslaw, K. S. and Curtius, J.: Ion-induced nucleation of pure biogenic particles, Nature, 533(7604), 521–526, doi:10.1038/nature17953, 2016.

Kuang, C., McMurry, P. H., McCormick, A. V. and Eisele, F. L.: Dependence of nucleation rates on sulfuric acid vapor concentration in diverse atmospheric locations, J. Geophys. Res., 113(D10), D10209, doi:10.1029/2007JD009253, 2008.

Kulmala, M., Lehtinen, K. E. J. and Laaksonen, A.: Cluster activation theory as an explanation of the linear dependence between formation rate of 3nm particles and sulphuric acid concentration, Atmos. Chem. Phys., 6(3), 787–793, doi:10.5194/acp-6-787-2006, 2006.

Kürten, A., Bianchi, F., Almeida, J., Kupiainen-Määttä, O., Dunne, E. M., Duplissy, J., Williamson, C., Barmet, P., Breitenlechner, M., Dommen, J., Donahue, N. M., Flagan, R. C., Franchin, A., Gordon, H., Hakala, J., Hansel, A., Heinritzi, M., Ickes, L., Jokinen, T., Kangasluoma, J., Kim, J., Kirkby, J., Kupc, A., Lehtipalo, K., Leiminger, M., Makhmutov, V., Onnela, A., Ortega, I. K., Petäjä, T., Praplan, A. P., Riccobono, F., Rissanen, M. P., Rondo, L., Schnitzhofer, R., Schobesberger, S., Smith, J. N., Steiner, G., Stozhkov, Y., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Wagner, P. E., Wimmer, D., Ye, P., Baltensperger, U., Carslaw, K., Kulmala, M. and Curtius, J.: Experimental particle formation rates spanning tropospheric sulfuric acid and ammonia abundances, ion production rates, and temperatures, J. Geophys. Res., 121(20), 12,377-12,400, doi:10.1002/2015JD023908, 2016.

Mandel, J.: Fitting Straight Lines When Both Variables are Subject to Error, J. Qual. Technol., 16(1), 1–14, doi:10.1080/00224065.1984.11978881, 1984.

Metzger, A., Verheggen, B., Dommen, J., Duplissy, J., Prevot, A. S. H., Weingartner, E., Riipinen, I., Kulmala, M., Spracklen, D. V, Carslaw, K. S. and Baltensperger, U.: Evidence for the role of organics in aerosol particle formation under atmospheric conditions., Proc. Natl. Acad. Sci. U. S. A., 107(15), 6646–51, doi:10.1073/pnas.0911330107, 2010.

Mikkonen, S., Romakkaniemi, S., Smith, J. N., Korhonen, H., Petäjä, T., Plass-Duelmer, C., Boy, M., McMurry, P. H., Lehtinen, K. E. J., Joutsensaari, J., Hamed, A., Mauldin III, R. L., Birmili, W., Spindler, G., Arnold, F., Kulmala, M. and Laaksonen, A.: A statistical proxy for sulphuric acid concentration, Atmos. Chem. Phys., 11(21), 11319–11334, doi:10.5194/acp-11-11319-2011, 2011.

Paasonen, P., Nieminen, T., Asmi, E., Manninen, H. E., Petäjä, T., Plass-Dülmer, C., Flentje, H., Birmili, W., Wiedensohler, A., Hõrrak, U., Metzger, A., Hamed, A., Laaksonen, A., Facchini, M. C., Kerminen, V. M. and Kulmala, M.: On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation, Atmos. Chem. Phys., 10(22), 11223–11242, doi:10.5194/acp-10-11223-2010, 2010.

Pitkänen, M. R. A., Mikkonen, S., Lehtinen, K. E. J., Lipponen, A. and Arola, A.: Artificial bias typically neglected in comparisons of uncertain atmospheric data, Geophys. Res. Lett., 43(18), 10,003-10,011, doi:10.1002/2016GL070852, 2016.

R Core Team: R: A language and environment for statistical computing., [online] Available from: http://www.r-project.org, 2018.

Riccobono, F., Schobesberger, S., Scott, C. E., Dommen, J., Ortega, I. K., Rondo, L., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Downard, A., Dunne, E. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Hansel, A., Junninen, H., Kajos, M., Keskinen, H., Kupc, A., Kürten, A., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Nieminen, T., Onnela, A., Petäjä, T., Praplan, A. P., Santos, F. D., Schallhart, S., Seinfeld, J. H., Sipilä, M., Spracklen, D. V, Stozhkov, Y., Stratmann, F., Tomé, A., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Wimmer, D., Carslaw, K. S., Curtius, J., Donahue, N. M., Kirkby, J., Kulmala, M., Worsnop, D. R. and Baltensperger, U.: Oxidation products of biogenic emissions contribute to nucleation of atmospheric particles., Science, 344(6185), 717–21, doi:10.1126/science.1243527, 2014.

Riipinen, I., Sihto, S.-L., Kulmala, M., Arnold, F., Dal Maso, M., Birmili, W., Saarnio, K., Teinilä, K., Kerminen, V.-M., Laaksonen, A. and Lehtinen, K. E. J.: Connections between atmospheric sulphuric acid and new particle formation during QUEST III–IV campaigns in Heidelberg and Hyytiälä, Atmos. Chem. Phys., 7(8), 1899–1914, doi:10.5194/acp-7-1899-2007, 2007.

Schennach, S. M.: Estimation of Nonlinear Models with Measurement Error, Econometrica, 72(1), 33–75, doi:10.1111/j.1468-0262.2004.00477.x, 2004.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: From air pollution to climate change. [online] Available from: https://www.wiley.com/en-fi/Atmospheric+Chemistry+and+Physics:+From+Air+Pollution+to+Climate+Change,+3rd+Edition-p-9781118947401 (Accessed 26 September 2018), 2016.

Sihto, S.-L., Kulmala, M., Kerminen, V.-M., Dal Maso, M., Petäjä, T., Riipinen, I., Korhonen, H., Arnold, F., Janson, R., Boy, M., Laaksonen, A. and Lehtinen, K. E. J.: Atmospheric sulphuric acid and aerosol formation: implications from atmospheric measurements for nucleation and early growth mechanisms, Atmos. Chem. Phys., 6(12), 4079–4091, doi:10.5194/acp-6-4079-2006, 2006.

Spiess, A.: Orthogonal Nonlinear Least-Squares Regression in R, [online] Available from: https://cran.hafro.is/web/packages/onls/vignettes/onls.pdf (Accessed 17 July 2018), 2015.

Spracklen, D. V., Carslaw, K. S., Kulmala, M., Kerminen, V.-M., Mann, G. W. and Sihto, S.-L.: The contribution of boundary

layer nucleation events to total particle concentrations on regional and global scales, Atmos. Chem. Phys., 6(12), 5631–5648, doi:10.5194/acp-6-5631-2006, 2006.

Stan Development Team: PyStan: the Python interface to Stan, Version 2.17.1.0., [online] Available from: http://mc-stan.org, 2018.

Therneau, T.: deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression. R package version 1.4., [online] Available from: https://cran.r-project.org/package=deming, 2018.

Trefall, H. and Nordö, J.: On Systematic Errors in the Least Squares Regression Analysis, with Application to the Atmospheric Effects on the Cosmic Radiation, Tellus, 11(4), 467–477, doi:10.3402/tellusa.v11i4.9324, 1959.

Tröstl, J., Chuang, W. K., Gordon, H., Heinritzi, M., Yan, C., Molteni, U., Ahlm, L., Frege, C., Bianchi, F., Wagner, R., Simon, M., Lehtipalo, K., Williamson, C., Craven, J. S., Duplissy, J., Adamov, A., Almeida, J., Bernhammer, A.-K., Breitenlechner, M., Brilke, S., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Gysel, M., Hansel, A., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Keskinen, H., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lawler, M., Leiminger, M., Mathot, S., Möhler, O., Nieminen, T., Onnela, A., Petäjä, T., Piel, F. M., Miettinen, P., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Sipilä, M., Smith, J. N., Steiner, G., Tomè, A., Virtanen, A., Wagner, A. C., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Carslaw, K. S., Curtius, J., Dommen, J., Kirkby, J., Kulmala, M., Riipinen, I., Worsnop, D. R., Donahue, N. M. and Baltensperger, U.: The role of low-volatility organic compounds in initial particle growth in the atmosphere, Nature, 533(7604), 527–531, doi:10.1038/nature18271, 2016.

Vehkamäki, H.: Classical nucleation theory in multicomponent systems, Springer-Verlag, Berlin/Heidelberg., 2006.

Weber, R. J., Marti, J. J., McMurry, P. H., Eisele, F. L., Tanner, D. J. and Jefferson, A.: Measurements of new particle formation and ultrafine particle growth rates at a clean continental site, J. Geophys. Res. Atmos., 102(D4), 4375–4385, doi:10.1029/96JD03656, 1997.

Wu, C. and Yu, J. Z.: Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting, Atmos. Meas. Tech., 11(2), 1233–1250, doi:10.5194/amt-11-1233-2018, 2018.

York, D.: Least-sqares fitting of a straight line, Can. J. Phys., 44(5), 1079–1086, doi:10.1139/p66-090, 1966.

York, D., Evensen, N. M., Martínez, M. L. and De Basabe Delgado, J.: Unified equations for the slope, intercept, and standard errors of the best straight line, Am. J. Phys., 72(3), 367–375, doi:10.1119/1.1632486, 2004.

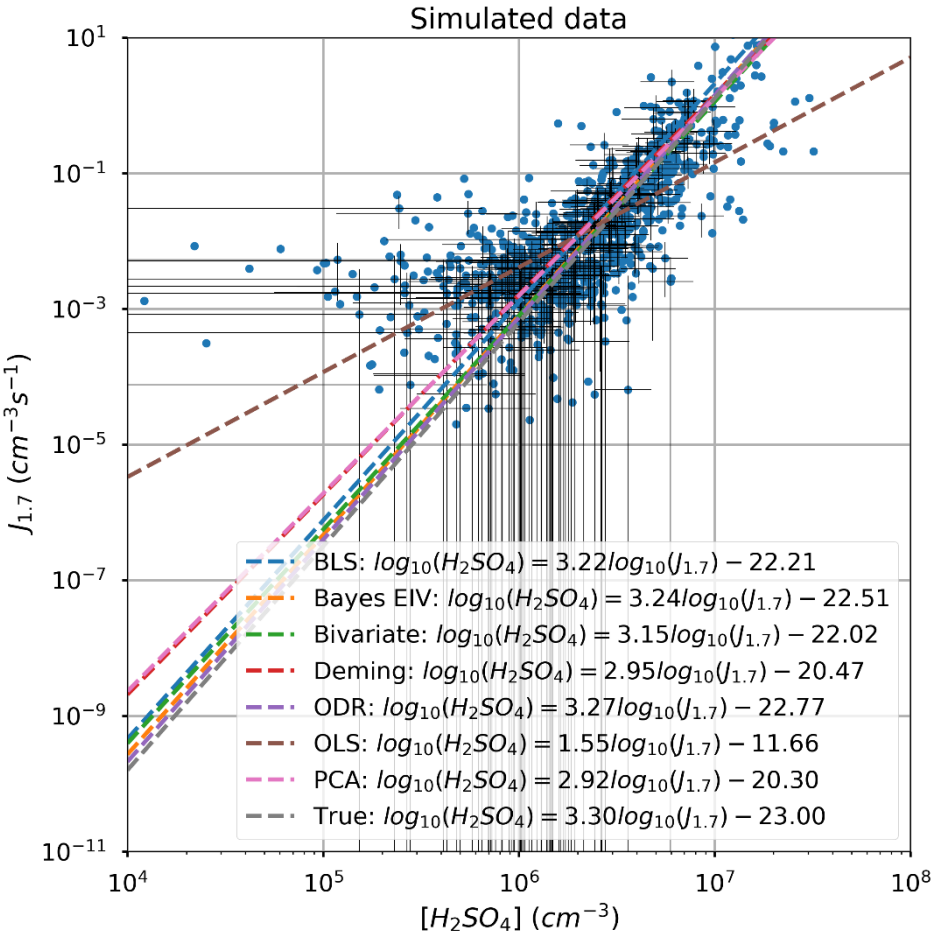

Figure 1. Regression lines fitted to the simulated data with all methods in comparison. Whiskers in data points refer to the measurement error used for simulation

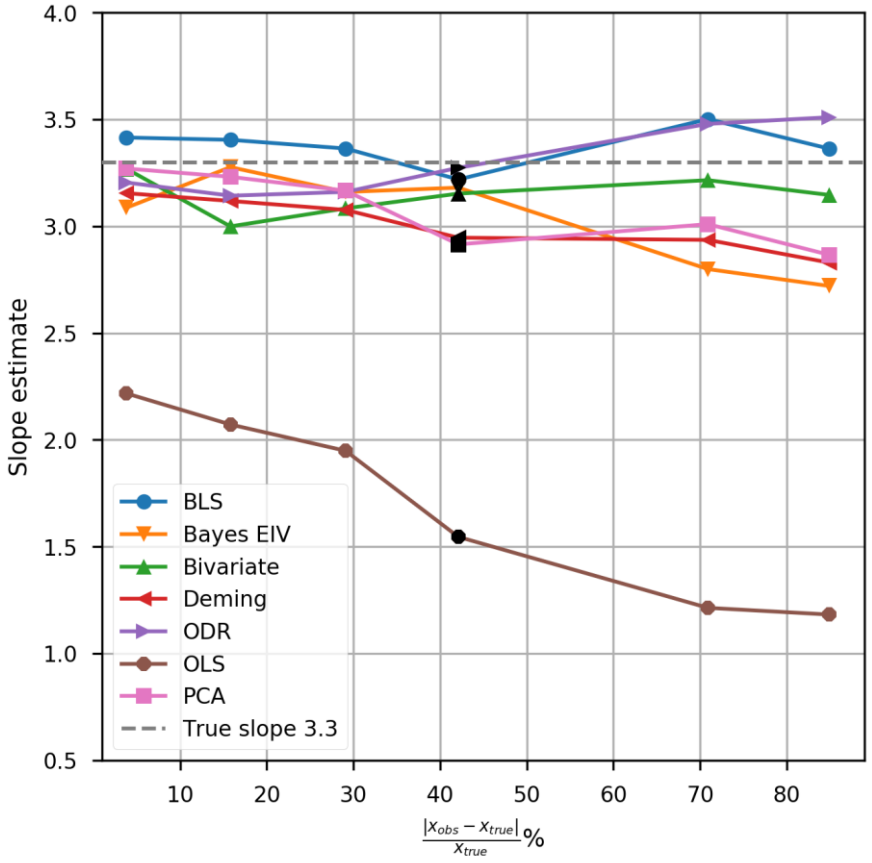

Figure 2. Sensitivity test for increasing uncertainty in simulated data. Black markers show the initial data set described in Section 3. Dashed line indicates the ~~"true slope"~~noise-free slope.

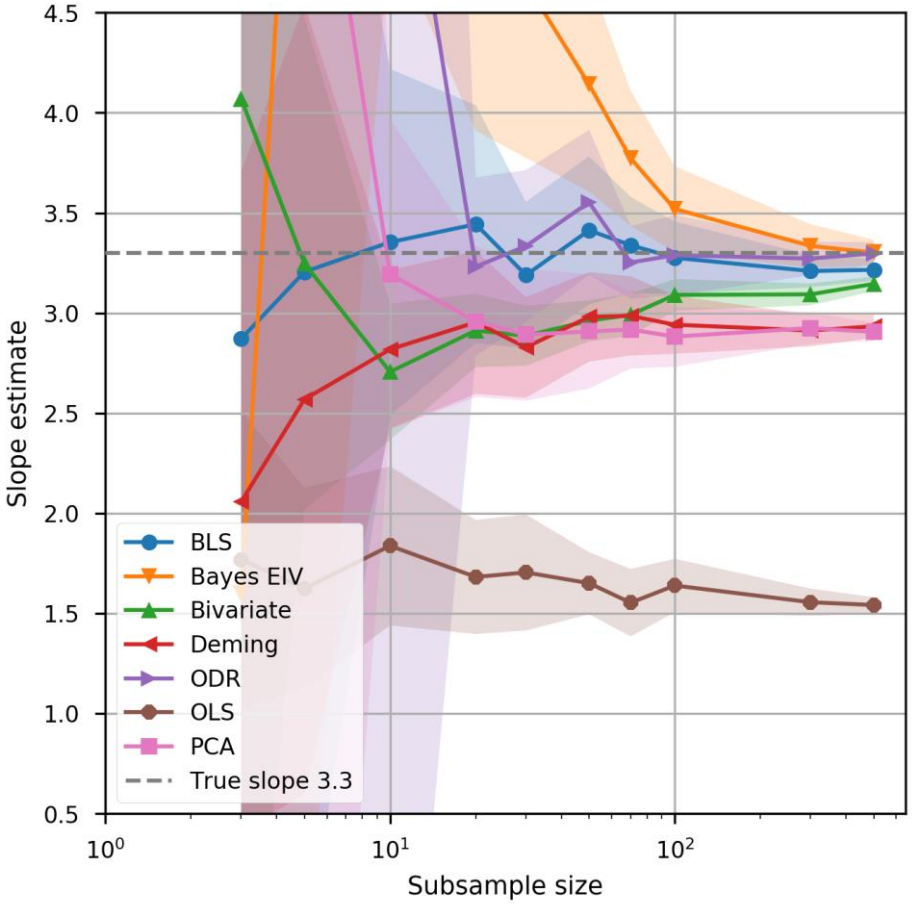

Figure 3. Effect of sample size on the uncertainty of different fits. Lines show the median and shading illustrates one standard deviation range of slope estimates for 40 repeated random samples. Dashed line indicates the ~~"true slope"~~noise-free slope.

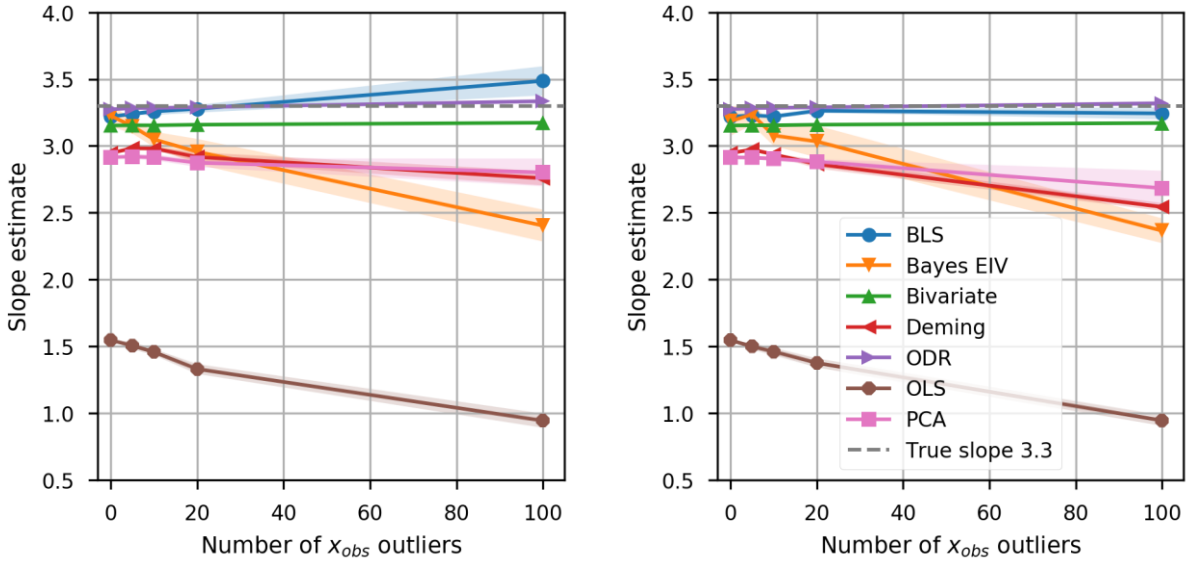

Figure 4. Effect of outliers in the data. Random outliers case on left panel and only high positives on right panel. Lines show the median and shading shows one standard deviation of slope estimates in ten repeated studies. Dashed line indicates the ~~"true slope"~~noise-free slope.

Table 1. The uncertainties used in simulation for sensitivity test for increasing uncertainty

dataset	σ_{abs}	σ_{rel}	Ratio (= (σ_{rel} * x'_{obs}) / σ_{abs})
1	10^3	0.05	315.0
2	10^4	0.18	113.4
3	$7*10^4$	0.3	27.0
4	$4*10^5$	0.3	4.7
5	$6.5*10^5$	0.45	4.4
6	10^6	0.55	3.5

Supplement S1 – Minimizing criteria for regression methods applied in the manuscript.

~~The minimizing criteria are given as follows:~~In this supplement, we introduce the minimizing criteria (C_{method}) for all methods applied in the main text. We also give the equations for regression coefficients ($\hat{\alpha}_{method}$ and $\hat{\beta}_{method}$) for the methods.

Ordinary Least Squares (OLS)

OLS minimizes the sum of squares vertical distances (residuals) between each point and the fitted line. OLS regression minimizes the following criterion:

$$C_{OLS} = \sum_{i=1}^{N}\left(y_i - \hat{\alpha}_{OLS} - \hat{\beta}_{OLS}x_i\right)^2 \quad (1)$$

where $\hat{\alpha}_{OLS}$ and $\hat{\beta}_{OLS}$ refer to estimators calculated from the data, given by

$$\hat{\beta}_{OLS} = \frac{S_x}{S_y},\ \hat{\alpha}_{OLS} = \bar{x} - \hat{\beta}_{OLS}\bar{y} \quad (2)$$

where observed variances for *x* $S_x = \sum_{i=1}^{N}(x_i - \bar{x})^2$ and~~,~~ for *y* $S_y = \sum_{i=1}^{N}(y_i - \bar{y})^2$, and observed covariance for *x* and *y* $S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

Orthogonal regression (ODR)

ODR (https://docs.scipy.org/doc/external/odrpack_guide.pdf, https://docs.scipy.org/doc/scipy/reference/odr.html, accessed 2018-07-27) minimizes the sum of the square of orthogonal distances between each point and the line, the criteria is given by

$$C_{ODR} = \sum_{i=1}^{N}\left(\left(x_i - \frac{y_i + x_i/\hat{\beta}_{ODR} - \hat{\alpha}_{ODR}}{\hat{\beta}_{ODR} + 1/\hat{\beta}_{ODR}}\right)^2 + \left(y_i - \hat{\alpha}_{ODR} - \frac{\hat{\beta}_{ODR}y_i + x_i - \hat{\alpha}_{ODR}\hat{\beta}_{ODR}}{\hat{\beta}_{ODR} + 1/\hat{\beta}_{ODR}}\right)^2\right) \quad (3)$$

Where

$$\hat{\beta}_{ODR} = \frac{S_y - S_x + \sqrt{(S_y - S_x)^2 + 4S_{xy}^2}}{2S_{xy}} \quad (4)$$

and

$$\hat{\alpha}_{ODR} = \bar{y} - \hat{\beta}_{ODR}\bar{x} \quad (5)$$

ODR takes into account ~~the~~ that errors exist in both axes but not the exact values of the~~ir~~ variances of variables. Thus only the ratio between the two error variances (λ_{xy}) is needed to improve the methodology. With notation of Francq and Govaerts (2014) this ratio is given by,

$$\lambda_{xy} = \frac{\sigma_y^2}{\sigma_x^2} \qquad (6)$$

where the numerator of the ratio is the error variance in the data in Y-axis and the denominator is the error variance in the data in X-axis.

Deming Regression (DR)

~~The **Deming Regression (**~~**DR**~~**)**~~ is the ML (Maximum Likelihood) solution of Eq. 1 when λ_{xy} is known. In practice, λ_{xy} is unknown and it is estimated from the variances of *x* and *y* calculated from the data.

The DR minimizes the criterion C_{DR} the sum of the square of (weighted) oblique distances between each point to the line

$$C_{DR} = \sum_{i=1}^{N}\left(\lambda_{xy}\left(x_i - \frac{y_i+\lambda_{xy}x_i/\hat{\beta}_{DR}-\hat{\alpha}_{DR}}{\hat{\beta}_{DR}+\lambda_{xy}/\hat{\beta}_{DR}}\right)^2 + \left(y_i - \hat{\alpha}_{DR} - \frac{\hat{\beta}_{DR}y_i+\lambda_{xy}x_i-\hat{\alpha}_{DR}\hat{\beta}_{DR}}{\hat{\beta}_{DR}+\lambda_{xy}/\hat{\beta}_{DR}}\right)^2\right) \qquad (7)$$

where

$$\hat{\beta}_{DR} = \frac{s_y-\lambda_{xy}s_x+\sqrt{(s_y-\lambda_{xy}s_x)^2+4\lambda_{xy}s_{xy}^2}}{2s_{xy}} \qquad (8)$$

and

$$\hat{\alpha}_{DR} = \bar{y} - \hat{\beta}_{DR}\bar{x} \qquad (9)$$

Bivariate Least Square regression, BLS

~~Bivariate Least Square regression,~~ **BLS**~~,~~ is a generic name but here we refer to the formulation described in Francq and Govaerts (2014) and references therein. BLS takes into account errors and heteroscedasticity in both axes and is written usually in matrix notation. BLS minimizes the criterion C_{BLS}, the sum of weighted residuals W_{BLS} given by:

$$C_{BLS} = \frac{1}{W_{BLS}}\sum_{i=1}^{N}\left(y_i - \hat{\alpha}_{BLS} - \hat{\beta}_{BLS}x_i\right)^2 \qquad (10)$$

with

$$W_{BLS} = \sigma_\varepsilon^2 = \frac{\sigma_y^2}{n_y} + \hat{\beta}_{BLS}^2\frac{\sigma_x^2}{n_x} \qquad (11)$$

Estimators for the parameters are computed by iterations with the following formulas:

$$\frac{1}{W_{BLS}}\begin{pmatrix} N & \sum_{i=1}^{N}x_i \\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i^2 \end{pmatrix}\begin{pmatrix}\hat{\alpha}_{BLS}\\ \hat{\beta}_{BLS}\end{pmatrix} = \frac{1}{W_{BLS}}\left(\sum_{i=1}^{N}\left(x_iy_i + \hat{\beta}_{BLS}\frac{\sigma_x^2}{n_x}\frac{\sum_{i=1}^{N}(y_i-\hat{\alpha}_{BLS}-\hat{\beta}_{BLS}x_i)^2}{W_{BLS}}\right)\right) \qquad (12)$$

Where known uncertainties $\sigma_x^2$and σ_y^2 are in this study replaced with estimated variances S_x and S_y.

A ~~S~~second Bivariate regression method that was used in this study is an implementation of the regression method described by **York** *et al.* (2004, Section III). The minimisation criterion is described in York *et al.* (1968) (York, D., 1968, 'Least squares fitting of a straight line with correlated errors', Earth and Planetary Science Letters (1969), pp. 320-324, North-Holland Publishing Company, Amsterdam.): ~~See their description of the method for details.~~

$$C_{york} = \sum_{i=0}^{N} \frac{1}{1-r_i^2}\left\{w(x_i)(x_{i,adj} - x_i)^2 - 2r\sqrt{w(x_i)w(y_i)}(x_{i,adj} - x_i)(y_{i,adj} - y_i) + w(y_i)(y_{i,adj} - y_i)^2\right\} \quad (13)$$

Where $w(x_i) = 1/\sigma_x^2$ and $w = (y_i)1/\sigma_y^2$ are the weight coefficients for x and y, respectively, and r is the correlation coefficient between x and y. $x_{i,adj}$ and $y_{i,adj}$ are adjusted values of x_i, y_i,that fulfill the requirement

$$y_{i,adj} = \hat{\alpha}_{york} + \hat{\beta}_{york} x_{i,adj} \quad (14)$$

The solution for $\hat{\alpha}_{york}$ and $\hat{\beta}_{york}$ is found iteratively following the ten step algorithm presented in **York** *et al.* (2004, Section III).

The Principal Component Analysis based regression (PCA)

PCA can be applied for bivariate and multivariate cases.

For one independent and one dependent variable, the regression line is

$y = \hat{\alpha}_{PCA} + \hat{\beta}_{PCA}x$ where the error between the *observed* value y_i and *estimated* value $a+bx_i$ is minimum. For n points data, we compute a and b by using the method of least squares that minimizes:

$$C_{PCA} = \sum_{i=1}^{N}\left(y_i - \hat{\alpha}_{PCA} - \hat{\beta}_{PCA}x_i\right)^2 \quad (15)$$

This is a standard technique that gives regression coefficients α and β.

$$\begin{bmatrix}\hat{\alpha}_{PCA} \\ \hat{\beta}_{PCA}\end{bmatrix} = \frac{\begin{bmatrix}S_x & -\bar{x} \\ -\bar{x} & 1\end{bmatrix}}{S_x - \bar{x}^2}\begin{bmatrix}\bar{y} \\ S_{xy}\end{bmatrix} \quad (16)$$

Bayesian error-in-variables regression (Bayes EIV)

Bayes EIV regression estimate applies Bayesian inference using the popular Stan software tool (http://mc-stan.org/users/documentation/, accessed 2018-07-27), which allowed the use of prior information of the model parameters. We assumed

β_{0BEIV} ~ student_t(5, 0.0, 100.0)
$\alpha_{BEIV}\beta_1$ ~ student_t(5, 0.0, 100.0)
x_{true} ~ lognormal(μ_x, σ_x)
$y_{true} = 10.0^{(\alpha_{BEIV} + \beta_{BEIV} * \log_{10}(x_{true}))}$,

where μ_x, and σ_x are the mean and standard deviation of x_{true} and y_{true} and are treated as unknowns. The observations x_{obs} and y_{obs} of x_{true} and y_{true}, respectively, were defined as:

x_{obs} ~ normal(x_{true}, $\sigma_{rel,x}$*x_{true} + $\sigma_{abs,x}$);
y_{obs} ~ normal(y_{true}, $\sigma_{rel,y}$*y_{true} + $\sigma_{abs,y}$);

where σ_{rel} and σ_{abs} are the relative and absolute components of standard uncertainties, respectively.

The Stan tool solved regression problems using 1000 iterations, and it provided a posteriori distributions for the model parameters β_{0BEIV} and $\alpha_{BEIV}\beta_1$. For the definitions of given student_t, lognormal, and normal probability distributions, see Stan documentation. In our regression analysis, we used the maximum a posteriori estimates for β_{BEIV} and $\alpha_{BEIV}\beta_0$ and β_1 provided by the software tool.