Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting

Cheng Wu\textsuperscript{1,2} and Jian Zhen Yu\textsuperscript{3,4,5}

\textsuperscript{1}Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China

\textsuperscript{2}Guangdong Provincial Engineering Research Center for on-line source apportionment system of air pollution, Guangzhou 510632, China

\textsuperscript{3}Division of Environment, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

\textsuperscript{4}Atmospheric Research Centre, Fok Ying Tung Graduate School, Hong Kong University of Science and Technology, Nansha, China

\textsuperscript{5}Department of Chemistry, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Corresponding to: Cheng Wu (wucheng.vip@foxmail.com) and Jian Zhen Yu (jian.yu@ust.hk)
Abstract

Linear regression techniques are widely used in atmospheric science, but are often improperly applied due to lack of consideration or inappropriate handling of measurement uncertainty. In this work, numerical experiments are performed to evaluate the performance of five linear regression techniques, significantly extending previous works by Chu and Saylor. The five techniques are Ordinary Least Square (OLS), Deming Regression (DR), Orthogonal Distance Regression (ODR), Weighted ODR (WODR), and York regression (YR). We first introduce a new data generation scheme that employs the Mersenne Twister (MT) pseudorandom number generator. The numerical simulations are also improved by: (a) refining the parameterization of non-linear measurement uncertainties, (b) inclusion of a linear measurement uncertainty, (c) inclusion of WODR for comparison. Results show that DR, WODR and YR produce an accurate slope, but the intercept by WODR and YR is overestimated and the degree of bias is more pronounced with a low $R^2_{XY}$ dataset. The importance of a properly weighting parameter $\lambda$ in DR is investigated by sensitivity tests, and it is found that an improper $\lambda$ in DR can lead to a bias in both the slope and intercept estimation. Because the $\lambda$ calculation depends on the actual form of the measurement error, it is essential to determine the exact form of measurement error in the XY data during the measurement stage. If a priori error in one of the variables is unknown, or the measurement error described cannot be trusted, DR, WODR and YR can provide the least biases in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.
1 Introduction

Linear regression is heavily used in atmospheric science to derive the slope and intercept of XY datasets. Examples of linear regression applications include primary OC (organic carbon) and EC (elemental carbon) ratio estimation (Turpin and Huntzicker, 1995; Lin et al., 2009), MAE (mass absorption efficiency) estimation from light absorption and EC mass (Moosmüller et al., 1998), source apportionment of polycyclic aromatic hydrocarbons using CO and NOx as combustion tracers (Lim et al., 1999), gas-phase reaction rate determination (Brauers and Finlayson-Pitts, 1997), inter-instrument comparison (Bauer et al., 2009; Cross et al., 2010; von Bobrutzki et al., 2010; Zieger et al., 2011; Wu et al., 2012; Huang et al., 2014; Zhou et al., 2016), inter-species analysis (Yu et al., 2005; Kuang et al., 2015), analytical protocol comparison (Chow et al., 2001; Chow et al., 2004; Cheng et al., 2011; Wu et al., 2016), light extinction budget reconstruction (Malm et al., 1994; Watson, 2002; Li et al., 2017), comparison between modeling and measurement (Petäjä et al., 2009), emission factor study (Janhäll et al., 2010), retrieval of shortwave cloud forcing (Cess et al., 1995), calculation of pollutant growth rate (Richter et al., 2005), estimation of ground level PM2.5 from MODIS data (Wang and Christopher, 2003), distinguishing OC origin from biomass burning using K+ as a tracer (Duan et al., 2004) and emission type identification by the EC/CO ratio (Chen et al., 2001).

Ordinary least squares (OLS) regression is the most widely used method due to its simplicity. In OLS, it is assumed that independent variables are error free. This is the case for certain applications, such as determining a calibration curve of an instrument in analytical chemistry. For example, a known amount of analyte (e.g., through weighing) can be used to calibrate the instrument output response (e.g., voltage). However, in many other applications, such as inter-instrument comparison, X and Y (from two instruments) may have comparable degrees of uncertainty. This deviation from the underlying assumption in OLS would produce biased slope and intercept when OLS is applied to the dataset.

To overcome the drawback of OLS, a number of error-in-variable regression models (also known as bivariate fittings (Cantrell, 2008) or total least-squares methods (Markovsky and Van Huffel, 2007) arise. Deming (1943) proposed an approach by
minimizing sum of squares of X and Y residuals. A closed-form solution of Deming regression (DR) was provided by York (1966). Method comparison work of various regression techniques by Cornbleet and Gochman (1979) found significant error in OLS slope estimation when the relative standard deviation (RSD) of measurement error in “X” exceeded 20%, while DR was found to reach a more accurate slope estimation. In an early application of the EC tracer method, Turpin and Huntzicker (1995) realized the limitation of OLS since OC and EC have comparable measurement uncertainty, thus recommended the use of DR for (OC/EC)pr (primary OC to EC ratio) estimation. Ayers (2001) conducted a simple numerical experiment and concluded that reduced major axis regression (RMA) is more suitable for air quality data regression analysis. Linnet (1999) pointed out that when applying DR for inter-method (or inter-instrument) comparison, special attention should be paid to the sample size. If the range ratio (max/min) is relatively small (e.g., less than 2), more samples are needed to obtain statistically significant results.

In principle, a best-fit regression line should have greater dependence on the more precise data points rather than the less reliable ones. Chu (2005) performed a comparison study of OLS and DR specifically focusing on the EC tracer method application, and found the slope estimated by DR is closer to the correct value than OLS but may still overestimate the ideal value. Saylor et al. (2006) extended the comparison work of Chu (2005) by including a regression technique developed by York et al. (2004). They found that the slope overestimation by DR in the study of Chu (2005) was due to improper configuration of the weighting parameter, λ. This λ value is the key to handling the uneven errors between data points for the best-fit line calculation. This example demonstrates the importance of appropriate weighting in the calculation of best-fit line for error-in-variable regression model, which is overlooked in many studies.

In this study, we extend the work by Saylor et al. (2006) to achieve four objectives. The first is to propose a new data generation scheme by applying the Mersenne Twister (MT) pseudorandom number generator for evaluation of linear regression techniques. In the study of Chu (2005), data generation is achieved by a varietal sine function, which has limitations in sample size, sample distribution, and nonadjustable correlation (R²) between X and Y. In comparison, the MT data generation provides more
flexibility, permitting adjustable sample size, XY correlation and distribution. The second is to develop a non-linear measurement error parameterization scheme for use in the regression method. The third is to incorporate linear measurement errors in the regression methods. In the work by Chu (2005) and Saylor et al. (2006), the relative measurement uncertainty ($\gamma_{unc}$) is non-linear with concentration, but a constant $\gamma_{unc}$ is often applied on atmospheric instruments due to its simplicity. The fourth is to include weighted orthogonal distance regression (WODR) for comparison.

Abbreviations and symbols used in this study are summarized in Table 1 for quick reference.

2 Description of regression techniques compared in this study

**Ordinary least squares (OLS) method.** OLS only considers the errors in dependent variables (Y). OLS regression is achieved by minimizing the sum of squares (S) in the Y residuals (i.e., distance of AB in Fig. S1):

$$S = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$  \hspace{1cm} (1)

where $Y_i$ are observed Y data points while $y_i$ are regressed Y data points of the regression line. N represents the number of data points that used for regression.

**Orthogonal distance regression (ODR).** ODR minimizes the sum of the squared orthogonal distances from all data points to the regressed line and considers equal error variances (i.e., distance of AC in Fig. S1):

$$S = \sum_{i=1}^{N}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]$$  \hspace{1cm} (2)

**Weighted orthogonal distance regression (WODR).** Unlike ODR that considers even error in X and Y, weightings based on measurement errors in both X and Y are considered in WODR when minimizing the sum of squared orthogonal distance from the data points to the regression line (Carroll and Ruppert, 1996) as shown by AD in Fig. S1:

$$S = \sum_{i=1}^{N}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2/\eta]$$  \hspace{1cm} (3)

where $\eta$ is error variance ratio that determines the angle $\theta$ shown in Fig. S1.

Implementation of ODR and WODR in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) was done by the computer routine ODRPACK95 (Boggs et al., 1989; Zwolak et al., 2007).
Deming regression (DR). Deming (1943) proposed the following function to minimize both the X and Y residuals as shown by AD in Fig. S1,
\[ S = \sum_{i=1}^{N} [\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2] \]  
(4)

where \( X_i \) and \( Y_i \) are observed data points and \( x_i \) and \( y_i \) are regressed data points. Individual data points are weighted based on errors in \( X_i \) and \( Y_i \),
\[ \omega(X_i) = \frac{1}{\sigma_{X_i}^2}, \quad \omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} \]  
(5)

where \( \sigma_{X_i} \) and \( \sigma_{Y_i} \) are the standard deviation of the error in measurement of \( X_i \) and \( Y_i \), respectively. The closed form solutions for slope and intercept of DR are shown in Appendix A.

York regression (YR). The York method (York et al., 2004) introduces the correlation coefficient of errors in X and Y into the minimization function.
\[ S = \sum_{i=1}^{N} \left[ \omega(X_i)(x_i - X_i)^2 - 2r_i\sqrt{\omega(X_i)\omega(Y_i)}(x_i - X_i)(y_i - Y_i) + \omega(Y_i)(y_i - Y_i)^2 \right] \frac{1}{1-r_i^2} \]  
(6)

where \( r_i \) is the correlation coefficient between measurement errors in \( X_i \) and \( Y_i \). The slope and intercept of YR are calculated iteratively through the formulas in Appendix A.

Summary of the five regression techniques is given in Table S1. It is worth noting that OLS and DR have closed-form expressions for calculating slope and intercept. In contrast, ODR, WODR and YR need to be solved iteratively. This need to be taken into consideration when choosing regression algorithm for handling huge amount of data.

A computer program (Scatter plot; Wu, 2017a) with graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed to facilitate the implementation of error-in-variables regression (including DR, WODR and YR). Two other Igor Pro based computer programs, Histbox (Wu, 2017b) and Aethalometer data processor (Wu, 2017c) are used for data analysis and visualization in this study.

3 Data description

Two types of data are used for regression comparison. The first type is synthetic data generated by computer programs, which can be used in the EC tracer method (Turpin
and Huntzicker, 1995) to demonstrate the regression application. The true “slope” and “intercept” are assigned during data generation, allowing quantitative comparison of the bias of each regression scheme. The second type of data comes from ambient measurement of light absorption, OC and EC in Guangzhou for demonstration in a real-world application.

3.1 Synthetic XY data generation

In this study, numerical simulations are conducted in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) through custom codes. Two types of generation schemes are employed, one is based on the MT pseudorandom number generator (Matsumoto and Nishimura, 1998) and the other is based on the sine function described by Chu (2005).

The general form of linear regression on XY data can be written as:

\[ Y = kX + b \]  

(7)

Here \( k \) is the regressed slope and \( b \) is the intercept. The underlying meaning is that, \( Y \) can be decomposed into two parts. One part is correlated with \( X \), and the ratio is defined by \( k \). The other part of \( Y \) is constant and independent of \( X \) and regarded as \( b \).

To make the discussion easier to follow, we intentionally avoid discussion using the abstract general form and instead opt to use a real-world application case in atmospheric science. Linear regression had been heavily applied on OC and EC data, here we use OC and EC data as an example to demonstrate the regression application in atmospheric science. In the EC tracer method, OC (mixture) is \( Y \) and EC (tracer) is \( X \). OC can be decomposed into three components based on their formation pathway:

\[ OC = POC_{comb} + POC_{non-comb} + SOC \]  

(8)

Here \( POC_{comb} \) is primary OC from combustion. \( POC_{non-comb} \) is primary OC emitted from non-combustion activities. \( SOC \) is secondary OC formed during atmospheric aging. Since \( POC_{comb} \) is co-emitted with EC and well correlated with each other, their relationship can be parameterized as:

\[ POC_{comb} = (OC/EC)_{pri} \times EC \]  

(9)

By carefully selecting an OC and EC subset when SOC is very low (considered as approximately zero), the combination of Eqs. (8) & (9) become:
\[ POC = (OC/EC)_{pri} \times EC + POC_{non-comb} \] (10)

The regressed slope of POC (Y) against EC (X) represents \((OC/EC)_{pri}\) (k in Eq. (7)). The regressed intercept become \(POC_{non-comb}\) (b in Eq. (7)). With known \((OC/EC)_{pri}\) and \(POC_{non-comb}\), SOC can be estimated by:

\[ SOC = OC - ((OC/EC)_{pri} \times EC + POC_{non-comb}) \] (11)

The data generation starts from EC (X values). Once EC is generated, \(POC_{comb}\) (the part of Y that is correlated with X) can be obtained by multiplying EC with a preset constant, \((OC/EC)_{pri}\) (slope k). Then the other preset constant \(POC_{non-comb}\) is added to \(POC_{comb}\) and the sum becomes POC (Y values). To simulate the real-world situation, measurement errors are added on X and Y values. Details of synthesized measurement error are discussed in the next section. Implementation of data generation by two types of mathematical schemes is explained in sect. 3.1.2 and 3.1.3, respectively.

### 3.1.1 Parameterization of synthesized measurement uncertainty

Weighting of variables is a crucial input for errors-in-variables linear regression methods such as DR, YR and WODR. In practice, the weights are usually defined as the inverse of the measurement error variance (Eq. (5)). When measurement errors are considered, measured concentrations (\(Conc_{measured}\)) are simulated by adding measurement uncertainties (\(\varepsilon_{Conc}\)) to the true concentrations (\(Conc_{true}\)):

\[ Conc_{measured} = Conc_{true} + \varepsilon_{Conc}. \] (12)

Here \(\varepsilon_{Conc}\) is the random error following an even distribution with an average of 0, the range of which is constrained by:

\[-\gamma_{Unc} \times Conc_{true} \leq \varepsilon_{Conc} \leq +\gamma_{Unc} \times Conc_{true} \] (13)

The \(\gamma_{Unc}\) is a dimensionless factor that describes the fractional measurement uncertainty relative to the true concentration (\(Conc_{true}\)). \(\gamma_{Unc}\) could be a function of \(Conc_{true}\) (Thompson, 1988) or a constant. The term \(\gamma_{Unc} \times Conc_{true}\) defines the boundary of random measurement errors.

Two types of measurement error are considered in this study. The first type is \(\gamma_{Unc-linear}\). In the data generation scheme of Chu (2005) for the measurement uncertainties (\(\varepsilon_{POC}\) and \(\varepsilon_{EC}\), \(\gamma_{Unc-linear}\) is non-linearly related to \(Conc_{true}\):

8
\[ Y_{\text{unc-linear}} = \frac{1}{\sqrt{\text{Conc}_{\text{true}}}} \]  

(14)

then Eq. (13) for POC and EC become:

\[ -\frac{1}{\sqrt{\text{POC}_{\text{true}}}} \times \text{POC}_{\text{true}} \leq \varepsilon_{\text{POC}} \leq +\frac{1}{\sqrt{\text{POC}_{\text{true}}}} \times \text{POC}_{\text{true}} \]

(15)

\[ -\frac{1}{\sqrt{\text{EC}_{\text{true}}}} \times \text{EC}_{\text{true}} \leq \varepsilon_{\text{EC}} \leq +\frac{1}{\sqrt{\text{EC}_{\text{true}}}} \times \text{EC}_{\text{true}} \]

(16)

In Eq. (14), the \( Y_{\text{unc}} \) decreases as concentration increases, since low concentrations are usually more challenging to measure. As a result, the \( Y_{\text{unc-linear}} \) defined in Eq. (14) is more realistic than the constant approach, but there are two limitations. First, the physical meaning of the uncertainty unit is lost. If the unit of OC is \( \mu g \text{ m}^{-3} \), then the unit of \( \varepsilon_{\text{OC}} \) becomes \( \sqrt{\mu g \text{ m}^{-3}} \). Second, the concentration is not normalized by a consistent relative value, making it sensitive to the X and Y units used. For example, if \( \text{POC}_{\text{true}}=0.9 \ \mu g \text{ m}^{-3} \), then \( \varepsilon_{\text{POC}} = \pm 0.95 \ \mu g \text{ m}^{-3} \) and \( Y_{\text{unc}} = 105\% \), but by changing the concentration unit to \( \text{POC}_{\text{true}}=900 \ \text{ng m}^{-3} \), then \( \varepsilon_{\text{OC}} = \pm 30 \ \text{ng m}^{-3} \) and \( Y_{\text{unc}} = 3\% \). To overcome these deficiencies, we propose to modify Eq. (14) to:

\[ Y_{\text{unc}} = \sqrt{\frac{\text{LOD}}{\text{Conc}_{\text{true}}}} \times \alpha \]  

(17)

here LOD (limit of detection) is introduced to generate a dimensionless \( Y_{\text{unc}} \). \( \alpha \) is a dimensionless adjustable factor to control the position of \( Y_{\text{unc}} \) curve on the concentration axis, which is indicated by the value of \( Y_{\text{unc}} \) at LOD level. As shown in Fig. 1a, at different values of \( \alpha \) (\( \alpha = 1, 0.5 \) and 0.3), the corresponding \( Y_{\text{unc}} \) at the same LOD level would be 100\%, 50\% and 30\%, respectively. By changing \( \alpha \), the location of the \( Y_{\text{unc}} \) curve on X axis direction can be set, using the \( Y_{\text{unc}} \) at LOD as the reference point. Then Eq. (17) for POC and EC become:

\[ -\frac{\text{LOD}_{\text{POC}}}{\sqrt{\text{POC}_{\text{true}}}} \times \alpha_{\text{POC}} \times \text{POC}_{\text{true}} \leq \varepsilon_{\text{POC}} \leq +\frac{\text{LOD}_{\text{POC}}}{\sqrt{\text{POC}_{\text{true}}}} \times \alpha_{\text{POC}} \times \text{POC}_{\text{true}} \]

(18)

\[ -\frac{\text{LOD}_{\text{EC}}}{\sqrt{\text{EC}_{\text{true}}}} \times \alpha_{\text{EC}} \times \text{EC}_{\text{true}} \leq \varepsilon_{\text{EC}} \leq +\frac{\text{LOD}_{\text{EC}}}{\sqrt{\text{EC}_{\text{true}}}} \times \alpha_{\text{EC}} \times \text{EC}_{\text{true}} \]

(19)

With the modified \( Y_{\text{unc-linear}} \) parameterization, concentrations of POC and EC are normalized by a corresponding LOD, which maintains unit consistency between
POC_{true} and \(\varepsilon_{POC}\) and EC_{true} and \(\varepsilon_{EC}\), and eliminates dependency on the concentration unit.

Uniform distribution has been used in previous studies (Cox et al., 2003; Chu, 2005; Saylor et al., 2006) and is adopted in this study to parameterize measurement error. For a uniform distribution in the interval \([a,b]\), the variance is \(\frac{1}{12} (a - b)^2\). Since \(\varepsilon_{POC}\) and \(\varepsilon_{EC}\) follow a uniform distribution in the interval as given by Eqs. (18) and (19), the weights in DR and YR (inverse of variance) become:

\[
\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{EC_{true} \times LOD_{EC} \times \sigma_{EC}^2}
\]

\[
\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{POC_{true} \times LOD_{POC} \times \sigma_{POC}^2}
\]

The parameter \(\lambda\) in Deming regression is then determined:

\[
\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{POC_{true} \times LOD_{POC} \times \sigma_{POC}^2}{EC_{true} \times LOD_{EC} \times \sigma_{EC}^2}
\]

Besides the \(\gamma_{Unc-\text{nonlinear}}\) discussed above, a second type measurement uncertainty parameterized by a constant proportional factor, \(\gamma_{Unc-\text{linear}}\), is very common in atmospheric applications:

\[
-\gamma_{POC_{unc}} \times POC_{true} \leq \varepsilon_{POC} \leq +\gamma_{POC_{unc}} \times POC_{true}
\]

\[
-\gamma_{EC_{unc}} \times EC_{true} \leq \varepsilon_{EC} \leq +\gamma_{EC_{unc}} \times EC_{true}
\]

where \(\gamma_{POC_{unc}}\) and \(\gamma_{EC_{unc}}\) are the relative measurement uncertainties, e.g., for relative measurement uncertainty of 10%, \(\gamma_{unc}=0.1\). As a result, the measurement error is linearly proportional to the concentration. An example comparison of \(\gamma_{Unc-\text{nonlinear}}\) and \(\gamma_{Unc-\text{linear}}\) is shown in Fig. 1b. For \(\gamma_{Unc-\text{linear}}\), the weights become:

\[
\omega(X_i) = \frac{1}{\sigma_{X_i}^2} = \frac{3}{(\gamma_{EC_{unc}} \times EC_{true})^2}
\]

\[
\omega(Y_i) = \frac{1}{\sigma_{Y_i}^2} = \frac{3}{(\gamma_{POC_{unc}} \times POC_{true})^2}
\]

and \(\lambda\) for Deming regression can be determined:

\[
\lambda = \frac{\omega(X_i)}{\omega(Y_i)} = \frac{(\gamma_{POC_{unc}} \times POC_{true})^2}{(\gamma_{EC_{unc}} \times EC_{true})^2}
\]
3.1.2 XY data generation by Mersenne Twister (MT) generator following a specific distribution

The Mersenne twister (MT) is a pseudorandom number generator (PRNG) developed by Matsumoto and Nishimura (1998). MT has been widely adopted by mainstream numerical analysis software (e.g., Matlab, SPSS, SAS and Igor Pro) as well as popular programing languages (e.g., R, Python, IDL, C++ and PHP). Data generation using MT provides a few advantages: (1) Frequency distribution can be easily assigned during the data generation process, allowing straightforward simulation of the frequency distribution characteristics (e.g., Gaussian or Log-normal) observed in ambient measurements; (2) The inputs for data generation are simply the mean and standard deviation of the data series and can be changed easily by the user; (3) The correlation \((R^2)\) between X and Y can be manipulated easily during the data generation to satisfy various purposes; (4) Unlike the sine function described by Chu (2005) that has a sample size limitation of 120, the sample size in MT data generation is highly flexible.

In this section, we will use POC as Y and EC as X as an example to explain the data generation. Procedure of applying MT to simulate ambient POC and EC data can be found in our previous study (Wu and Yu, 2016). Details of the data generation steps are shown in Fig. 2 and described below. The first step is generation of ECtrue by MT. In our previous study, it was found that ambient POC and EC data follow a lognormal distribution in various locations of the Pearl River Delta (PRD) region. Therefore, lognormal distributions are adopted during ECtrue generation. A range of average concentration and relative standard deviation (RSD) from ambient samples is considered in formulating the lognormal distribution. The second step is to generate POCcomb. As shown in Fig. 2, POCcomb is generated by multiplying ECtrue with \((OC/EC)_{pri}\). Instead of having a Gaussian distribution, \((OC/EC)_{pri}\) in this study is a single value, which favors direct comparison between the true value of \((OC/EC)_{pri}\) and \((OC/EC)_{pri}\) estimated from the regression slope. The third step is generation of POCtrue by adding POCnon-comb onto POCcomb. Instead of having a distribution, POCnon-comb in this study is a single value, which favors direct comparison between the true value of POCnon-comb and POCnon-comb estimated from the regression intercept. The fourth step is to compute \(\varepsilon_{POC}\) and \(\varepsilon_{EC}\). As discussed in sect. 3.1.1, two types of measurement errors are considered for \(\varepsilon_{POC}\) and \(\varepsilon_{EC}\) calculation: \(\gamma_{Unc\text{-}nonlinear}\) and \(\gamma_{Unc\text{-}linear}\). In the
last step, \( \text{POC}_{\text{measured}} \) and \( \text{EC}_{\text{measured}} \) are calculated following Eq. (12), i.e., applying measurement errors on \( \text{POC}_{\text{true}} \) and \( \text{EC}_{\text{true}} \). Then \( \text{POC}_{\text{measured}} \) and \( \text{EC}_{\text{measured}} \) can be used as \( Y \) and \( X \), respectively, to test the performance of various regression techniques. An Igor Pro based program with graphical user interface (GUI) is developed to facilitate the MT data generation for OC and EC. A brief introduction is given in the Supplemental Information.

**3.1.3 XY data generation by the sine function of Chu (2005)**

Beside MT, inclusion of the sine function data generation scheme in this study mainly serves two purposes. First, the sine function scheme was adopted in two previous studies (Chu, 2005; Saylor et al., 2006), the inclusion of this scheme can help to verify whether the codes in Igor for various regression approaches yield the same results from the two previous studies. Second, the crosscheck between results from sine function and MT provides circumstantial evidence that the MT scheme works as expected.

In this section, XY data generation by sine functions is demonstrated using POC as \( Y \) and EC as \( X \). There are four steps in POC and EC data generation as shown by the flowchart in Fig. S2. Details are explained as follows: (1) The first step is to generate POC and EC (Chu, 2005):

\[
\begin{align*}
\text{POC}_{\text{comb}} &= 14 + 12(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi)) \\
\text{EC}_{\text{true}} &= 3.5 + 3(\sin\left(\frac{x}{\tau}\right) + \sin(x - \phi))
\end{align*}
\]  

(28)  

(29)

Here \( x \) is the elapsed hour (\( x=1,2,3,\ldots,n; n\leq120 \)), \( \tau \) is used to adjust the width of each peak, and \( \phi \) is used to adjust the phase of the sine wave. The constants 14 and 3.5 are used to lift the sine wave to the positive range of the \( Y \) axis. An example of data generation by the sine functions of Chu (2005) is shown in Fig. 3. Dividing Eq. (28) by Eq. (29) yields a value of 4. In this way the exact relation between POC and EC is defined clearly as \( \text{(OC/EC)}_{\text{pri}} = 4 \). (2) With POC\(_{\text{comb}}\) and EC\(_{\text{true}}\) generated, the second step is to add POC\(_{\text{non-comb}}\) to POC\(_{\text{comb}}\) to compute POC\(_{\text{true}}\). As for POC\(_{\text{non-comb}}\), a single value is assigned and added to all POC following Eq. (10). Then the goodness of the regression intercept can be evaluated by comparing the regressed intercept with preset POC\(_{\text{non-comb}}\). (3) The third step is to compute \( \varepsilon_{\text{POC}} \) and \( \varepsilon_{\text{EC}} \), considering both \( \gamma_{\text{unc-nonlinear}} \) and \( \gamma_{\text{unc-linear}} \). (4) The last step is to apply measurement errors on...
POC\textsubscript{true} and EC\textsubscript{true} following Eq. (12). Then POC\textsubscript{measured} and EC\textsubscript{measured} can be used as Y and X, respectively, to evaluate the performance of various regression techniques.

### 3.2 Ambient measurement of $\sigma_{abs}$ and EC

Sampling was conducted from Feb 2012 to Jan 2013 at the suburban Nancun (NC) site (23° 0'11.82"N, 113°21'18.04"E), which is situated on top of the highest peak (141 m ASL) in the Panyu district of Guangzhou. This site is located at the geographic center of Pearl River Delta region (PRD), making it a good location for representing the average atmospheric mixing characteristics of city clusters in the PRD region. Light absorption measurements were performed by a 7λ Aethalometer (AE-31, Magee Scientific Company, Berkeley, CA, USA). EC mass concentrations were measured by a real time ECOC analyzer (Model RT-4, Sunset Laboratory Inc., Tigard, Oregon, USA). Both instruments utilized inlets with a 2.5 μm particle diameter cutoff. The algorithm of Weingartner et al. (2003) was adopted to correct the sampling artifacts (aerosol loading, filter matrix and scattering effect) (Coen et al., 2010) in Aethalometer measurement. A customized computer program with graphical user interface, Aethalometer data processor (Wu et al., 2018), was developed to perform the data correction and detailed descriptions can be found in https://sites.google.com/site/wuchengust. More details of the measurements can be found in Wu et al. (2018).

### 4 Comparison study using synthetic data

In the following comparisons, six regression approaches are compared using two data generation schemes (Chu sine function and MT) separately, as illustrated in Fig. 4. Each data generation scheme considers both $\gamma_{\text{unc¬linear}}$ and $\gamma_{\text{unc¬linear}}$ in measurement error parameterization. In total, 18 cases are tested with different combination of data generation schemes, measurement error parameterization schemes, true slope and intercept settings. In each case, six regression approaches are tested, including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(x_i)}{\omega(y_i)}$), ODR, WODR and YR. In commercial software (e.g., OriginPro®, SigmaPlot®, GraphPad Prism®, etc), $\lambda$ in DR is set to 1 by default if not specified. As indicated by Saylor et al. (2006), the bias observed in the study of Chu (2005) is likely due to $\lambda = 1$ in DR. The purpose of including DR ($\lambda = 1$) in this study...
is to examine the potential bias using the default input in many software products. The six regression approaches are considered to examine the sensitivity of regression results to various parameters used in data generation. For each case, 5000 runs are performed to obtain statistically significant results, as recommended by Saylor et al. (2006). The mean slope and intercept from 5000 runs is compared with the true value assigned during data generation. If the difference is <5%, the result is considered unbiased.

4.1 Comparison results using the data set of Chu (2005)

In this section, the scheme of Chu (2005) is adopted for data generation to obtain a benchmark of six regression approaches. With different setup of slope, intercept and \( y_{Unc} \), 6 cases (Case 1 ~ 6) are studied and the results are discussed below.

4.1.1 Results with \( y_{Unc\text{-nonlinear}} \)

A comparison of the regression techniques results with \( y_{Unc\text{-nonlinear}} \) (following Eqs. (18) & (19)) is summarized in Table 2. LOD_{POC}, LOD_{EC}, \( \alpha_{POC} \) and \( \alpha_{EC} \) are all set to 1 to reproduce the data studied by Chu (2005) and Saylor et al. (2006). Two sets of true slope and intercept are considered (Case 1: Slope=4, Intercept=0; Case 2: Slope=4, Intercept=3) to examine if any results are sensitive to the non-zero intercept. The \( R^2 \) (POC, EC) from 5000 runs for both case 1 and 2 are 0.67±0.03.

As shown in Fig. 5, for the zero-intercept case (Case 1), OLS significantly underestimates the slope (2.95±0.14) while overestimates the intercept (5.84±0.78). This result indicates that OLS is not suitable for errors-in-variables linear regression, consistent with similar analysis results from Chu (2005) and Saylor et al. (2006). With DR, if the \( \lambda \) is properly calculated by weights (\( \lambda = \frac{\sigma(X)}{\sigma(Y)} \)), unbiased slope (4.01±0.25) and intercept (-0.04±1.28) are obtained; however, results from DR with \( \lambda=1 \) show obvious bias in the slope (4.27±0.27) and intercept (-1.45±1.36). ODR also produces biased slope (4.27±0.27) and intercept (-1.45±1.36), which are identical to results of DR when \( \lambda=1 \). With WODR, unbiased slope (3.98±0.22) is observed, but the intercept is overestimated (1.12±1.02). Results of YR are identical to WODR. For Case 2 (slope=4, intercept=3), slopes from all six regression approaches are consistent with...
Case 1 (Table 2). The Case 2 intercepts are equal to the Case 1 intercepts plus 3, implying that all the regression methods are not sensitive to a non-zero intercept.

For case 3, $LOD_{POC} = 0.5$, $LOD_{EC} = 0.5$, $\alpha_{POC} = 0.5$, $\alpha_{EC} = 0.5$ are adopted (Table 2), leading to an offset to the left of $\gamma_{Unc - nonlinear}$ (blue curve) compared to Case 1 and 2 (black curve) in Fig. 1. As a result, for the same concentration of EC and OC in Case 3, the $\gamma_{Unc - nonlinear}$ is smaller than in Case 1 and Case 2 as indicated by a higher $R^2$ (0.95±0.01 for Case 3, Table 2). With a smaller measurement uncertainty, the degree of bias in Case 3 is smaller than in Case 1. For example, OLS slope is less biased in Case 3 (3.83±0.08) compared to Case 1 (2.94±0.14). Similarly, the slope (4.03±0.09) and intercept (-0.18±0.44) of DR ($\lambda = 1$) exhibit a much smaller bias with a smaller measurement uncertainty, implying that the degree of bias by improperly weighting in DR, WODR and YR is associated with the degree of measurement uncertainty. A higher measurement uncertainty results in larger bias in slope and intercept.

An uneven $LOD_{POC}$ and $LOD_{EC}$ is tested in Case 4 with $LOD_{POC} = 1$, $LOD_{EC} = 0.5$, $\alpha_{POC} = 0.5$, $\alpha_{EC} = 0.5$, which yield a $R^2$(POC, EC) of 0.78±0.02. The results are similar to Case 1. For DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) unbiased slope and intercept are obtained. For WODR and YR, unbiased slopes are reported with a small bias in the intercepts. Large bias values are observed in both the slopes and intercepts in Case 4 using OLS, DR ($\lambda = 1$) and ODR.

### 4.1.2 Results with $\gamma_{Unc - linear}$

Cases 5 and 6 represent the results from using $\gamma_{Unc - linear}$ and are shown in Table 2. $\gamma_{Unc}$ is set to 30% to achieve a $R^2$ (POC, EC) of 0.7, a value close to the $R^2$ in studies of Chu (2005) and Saylor et al. (2006). In Case 5 (slope=4, intercept=0), unbiased slopes and intercepts are determined by DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR. OLS underestimates the slope (3.32 ± 0.20) and overestimates intercept (3.77 ± 0.90), while DR ($\lambda = 1$) and ODR overestimate the slopes (4.75 ± 0.30) and underestimate the intercepts (-4.14 ±1.36). In Case 6 (slope=4, intercept=3), results similar to Case 5 are obtained. It is worth noting that although the mean intercept (3.05±1.22) of DR ($\lambda =$
\[
\frac{\omega(X_i)}{\omega(Y_i)},
\]
is closest to the true value (intercept=3), the deviations are much larger than for WODR (2.72±0.74).

4.2 Comparison results using data generated by MT

In this section, MT is adopted for data generation to obtain a benchmark of six regression approaches. Both \( \gamma_{Unc-linear} \) and \( \gamma_{Unc-linear} \) are considered. With different configuration of slope, intercept and \( \gamma_{Unc} \), 12 cases (Case 7 ~ Case 18) are studied and the results are discussed below.

4.2.1 \( \gamma_{Unc-linear} \) results

Cases 7 and 8 use data generated by MT and \( \gamma_{Unc-linear} \) with results shown in Table 2. In Case 7 (slope=4, intercept=0, \( LOD_{POC}=1 \), \( LOD_{EC}=1 \), \( \alpha_{POC}=1 \), \( \alpha_{EC}=1 \)), unbiased slope (4.00 ±0.03) and intercept (0.00 ±0.17) is estimated by DR (\( \lambda = \frac{\omega(X)}{\omega(Y)} \)). WODR and YR yield unbiased slopes (3.96 ±0.03) but overestimate the intercepts (1.21 ±0.13). DR (\( \lambda = 1 \)) and ODR report slightly biased slopes (4.17 ±0.04) with biased intercepts (-0.94 ±0.18). OLS underestimates the slope (3.22 ±0.03) and overestimates the intercept (4.30 ±0.14). In Case 8 (slope=4, intercept=3, \( LOD_{POC}=1 \), \( LOD_{EC}=1 \), \( \alpha_{POC}=1 \), \( \alpha_{EC}=1 \)), DR (\( \lambda = \frac{\omega(X)}{\omega(Y)} \)) provides unbiased slope (4.00 ±0.03) and intercept (3.00 ±0.18) estimations. WODR and YR report unbiased slopes (3.97 ±0.03) and overestimate intercepts (4.11 ±0.13). OLS, DR (\( \lambda = 1 \)) and ODR report biased slopes and intercepts.

To test the overestimation/underestimation dependency on the true slope, Case 9 (slope=0.5, intercept=0, \( LOD_{POC} =1 \), \( LOD_{EC} =1 \), \( \alpha_{POC} =1 \), \( \alpha_{EC} =1 \)) and case 10 (slope=0.5, intercept=3, \( LOD_{POC} =1 \), \( LOD_{EC} =1 \), \( \alpha_{POC} =1 \), \( \alpha_{EC} =1 \)) are conducted and the results are shown in Table 2. Unlike the overestimation observed in Case 1~Case 8, DR (\( \lambda = 1 \)) and ODR underestimate the slopes (0.46 ±0.01) in Case 9. In case 10, DR (\( \lambda = 1 \)), DR (\( \lambda = \frac{\omega(X)}{\omega(Y)} \)) and ODR report unbiased slopes and intercepts. Case 11 and case 12 test the bias when the true slope is 1 as shown in Table 2. In Case 11 (intercept=0), all regression approaches except OLS can provide unbiased results. In Case 12, all regression approaches report unbiased slopes except OLS, but DR (\( \lambda = \frac{\omega(X)}{\omega(Y)} \)) is the only regression approach that reports unbiased intercept.
These results imply that if the true slope is less than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to underestimate slope. If the true slope is 1, these two estimators can provide unbiased results. If the true slope is larger than 1, the improper weighting ($\lambda = 1$) in Deming regression and ODR without weighting tends to overestimate slope.

4.2.2 $\gamma_{unc\text{-}linear}$ results

Cases 13 and 14 (Table 2) represent the results from using $\gamma_{unc\text{-}linear}$ (30%) and data generated from MT. For case 13 (slope=4, intercept=0), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR provide the best estimation of slopes and intercepts. DR ($\lambda = 1$) and ODR overestimate slopes (4.53 ±0.05) and underestimate intercepts (-2.94 ±0.24). For case 14 (slope=4, intercept=3), DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$), WODR and YR provide an unbiased estimation of slopes. But DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) is the only regression approach reporting unbiased intercept (3.08 ±0.23). Cases 15 and 16 are tested to investigate whether the results are different if the true slope is smaller than 1. As shown in Table 2, the results are similar to case 13&14 that DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) can provide unbiased slope and intercept while WODR and YR can provide unbiased slopes but biased intercepts. Cases 17 and 18 are tested to see if the results are the same for a special case when the true slope is 1. As shown in Table 2, the results are similar to case 13&14, implying that these results are not sensitive to the special case when the true slope is 1.

4.3 The importance of appropriate $\lambda$ input for Deming regression

As discussed above, inappropriate $\lambda$ assignment in the Deming regression (e.g., $\lambda=1$ by default for much commercial software) leads to biased slope and intercept. Beside $\lambda = 1$, inappropriate $\lambda$ input due to improper handling of measurement uncertainty can also result in bias for Deming regression. An example is shown in Fig. S3. Data is generated by MT with following parameters: slope=4, intercept=0, and $\gamma_{unc\text{-}linear}$ (30%). Fig. S2 a&b demonstrates that when an appropriate $\lambda$ is provided (following $\gamma_{unc\text{-}linear}$, $\lambda = \frac{P \sigma^2}{E \sigma^2}$), unbiased slopes and intercepts are obtained. If an improper $\lambda$ is used due to a mismatched measurement uncertainty assumption ($\gamma_{unc\text{-}nonlinear}$, $\lambda = \frac{P \sigma_C^2}{E \sigma}$), the
slopes are overestimated (Fig. S3c, 4.37±0.05) and intercepts are underestimated (Fig. S3d, -2.01±0.24). This result emphasizes the importance of determining the correct form of measurement uncertainty in ambient samples, since λ is a crucial parameter in Deming regression.

In the λ calculation, different representations for POC and EC, including mean, median and mode, are tested as shown in Fig. S4. The results show that when X and Y have a similar distribution (e.g., both are log-normal), any of mean, median or mode can be used for the λ calculation.

4.4 Caveats of regressions with unknown X and Y uncertainties

In atmospheric applications, there are scenarios in which a priori error in one of the variables is unknown, or the measurement error described cannot be trusted. For example, in the case of comparing model prediction and measurement data, the uncertainty of model prediction data is unknown. A second example is the case in which measurement uncertainty cannot be determined due to the lack of duplicated or collocated measurements and as a result, an arbitrarily assumed uncertainty is used. Such a case was illustrated in the study by Flanagan et al. (2006). They found that in the Speciation Trends Network (STN), the whole-system uncertainty retrieved by data from collocated samplers was different from the arbitrarily assumed 5% uncertainty. Additionally, the discrepancy between the actual uncertainty obtained through collocated samplers and the arbitrarily assumed uncertainty varied by chemical species.

To investigate the performance of different regression approaches in these cases, two tests (A and B) are conducted.

In Test A, the actual measurement error for X is fixed at 30% while $\gamma_{unc}$ for Y varies from 1% to 50%. The assumed measurement error for regression is 10% for both X and Y. Result of Test A are shown in Figs. 6 a and b. For OLS, the slopes are underestimated (-14 ~ -12%) and intercepts are overestimated (90 ~ 103%) and the biases are independent of variations in $\gamma_{unc,Y}$. ODR and DR ($\lambda = 1$) yield similar results with over-estimated slopes (0 ~ 44%) and under-estimated intercepts (-330 ~ 0%). The degree of bias in slopes and intercepts depends on the $\gamma_{unc,Y}$. WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR
perform much better than other regression approaches in Test A, with a smaller bias in
both slopes (-8 ~ 12%) and intercepts -98 ~ 55%).

In Test B, $y_{unc,Y}$ is fixed at 30% and $y_{unc,X}$ varies between 1 ~ 50%. The results of Test
B are shown in Figs. 6 c and d. The assumed measurement error for regression is 10%
for both X and Y. OLS underestimates the slopes (-29 ~-0.2%) and overestimates the
intercepts (2 ~ 209%). In contrast to Test A in which slope and intercept biases are
independent of variations in $y_{unc,Y}$, the slope and intercept biases in Test B exhibit
dependency on $y_{unc,X}$. The reason behind is because OLS only considers errors in Y
and X is assumed to be error free. ODR and DR ($\lambda = 1$) yield similar results with over-
estimated slopes (11 ~ 18%) and under-estimated intercepts (-144 ~ -87%). The degree
of bias in slopes and intercepts is relatively independent on the $y_{unc,X}$. WODR, DR
($\lambda = \frac{\omega(X)}{\omega(Y)}$) and YR performed much better than the other regression approaches in Test
B, with a smaller bias in both slopes (-14 ~ 8%) and intercepts (-59 ~ 106%).

The results from these two tests suggest that, in case of one of the measurement error
described cannot be trusted or a priori error in one of the variables is unknown, WODR,
DR ($\lambda = \frac{\omega(X)}{\omega(Y)}$) and YR should be used instead of ODR and DR ($\lambda = 1$) and OLS. This
conclusion is consistent with results presented in sect. 4.1 and 4.2. This analysis, albeit
crude, also suggests that, in general, the magnitude of bias in slope estimation by these
regression approaches is smaller than those for intercept. In other words, slope is a more
reliable quantity compared to intercept when extracting quantitative information from
linear regressions.

5 Regression applications to ambient data

This section demonstrates the application of the 6 regression approaches on a light
absorption coefficient and EC dataset collected in a suburban site in Guangzhou. As
mentioned in sect. 4.4, measurement uncertainties are crucial inputs for DR, YR and
WODR. The measurement precision of Aethalometer is 5% (Hansen, 2005) while EC
by RT-ECOC analyzer is 24% (Bauer et al., 2009). These measurement uncertainties
are used in DR, YR and WODR calculation. The data-set contains 6926 data points
with a $R^2$ of 0.92.
As shown in Fig. 7, Y axis is light absorption at 520 nm ($\sigma_{abs520}$) and the X axis is EC mass concentration. The regressed slopes represent the mass absorption efficiency (MAE) of EC at 520 nm, ranging from 13.66 to 15.94 m$^2$g$^{-1}$ by the six regression approaches. OLS yields the lowest slope (13.66 as shown in Fig. 7a) among all six regression approaches, consistent with the results using synthetic data. This implies that OLS tends to underestimate regression slope when mean Y to X ratio is larger than 1. DR ($\lambda = 1$) and ODR report the same slope (14.88) and intercept (5.54), this equivalency is also observed for the synthetic data. Similarly, WODR and YR yield identical slope (14.88) and intercept (5.54), in line with the synthetic data results. The regressed slope by DR ($\lambda = 1$) is higher than DR ($\lambda = \frac{\omega(X)}{\omega(Y)}$), and this relationship agrees well with the synthetic data results.

Regression comparison is also performed on hourly OC and EC data. Regression on OC/EC percentile subset is a widely used empirical approach for primary OC/EC ratio determination. Fig. S5 shows the regression slopes as a function of OC/EC percentile. OC/EC percentile ranges from 0.5% to 100%, with an interval of 0.5%. As the percentile increases, SOC contribution in OC increases as well, resulting in decreased R$^2$ between OC and EC. The deviations between six regression approaches exhibit a dependency on R$^2$. When percentile is relatively small (e.g., <10%), the differences between the six regression approaches are also small due to the high R$^2$ (0.98). The deviations between the six regression approaches become more pronounced as R$^2$ decreases (e.g., <0.9). The deviations are expected to be even larger when R$^2$ is less than 0.8. These results emphasize the importance of applying error-in-variables regression, since ambient XY data more likely has a R$^2$ less than 0.9 in most cases.

As discussed in this section, the ambient data confirm the results obtained in comparing methods with the synthetic data. The advantage of using the synthetic data for regression approaches evaluation is that the ideal slope and intercept are known values during the data generation, so the bias of each regression approach can be quantified.

6 Recommendations and conclusions

This study aims to provide a benchmark of commonly used linear regression algorithms using a new data generation scheme (MT). Six regression approaches are tested,
including OLS, DR ($\lambda = 1$), DR ($\lambda = \frac{\omega(X)}{\omega(Y)}$), ODR, WODR and YR. The results show that OLS fails to estimate the correct slope and intercept when both X and Y have measurement errors. This result is consistent with previous studies. For ambient data with $R^2$ less than 0.9, error-in-variables regression is needed to minimize the biases in slope and intercept. If measurement uncertainties in X and Y are determined during the measurement, measurement uncertainties should be used for regression. With appropriate weighting, DR, WODR and YR can provide the best results among all tested regression techniques. Sensitivity tests also reveal the importance of the weighting parameter $\lambda$ in DR. An improper $\lambda$ could lead to biased slope and intercept. Since the $\lambda$ estimation depends on the form of the measurement errors, it is important to determine the measurement errors during the experimentation stage rather than making assumptions. If measurement errors are not available from the measurement and assumptions are made on measurement errors, DR, WODR and YR are still the best option that can provide the least bias in slope and intercept among all tested regression techniques. For these reasons, DR, WODR and YR are recommended for atmospheric studies when both X and Y data have measurement errors.

Application of error-in-variables regression is often overlooked in atmospheric studies, partly due to the lack of a specified tool for the regression implementation. To facilitate the implementation of error-in-variables regression (including DR, WODR and YR), a computer program (Scatter plot) with graphical user interface (GUI) in Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) is developed (Fig. 8). It is packed with many useful features for data analysis and plotting, including batch plotting, data masking via GUI, color coding in Z axis, data filtering and grouping by numerical values and strings. The Scatter plot program and user manual are available from [https://sites.google.com/site/wuchengust](https://sites.google.com/site/wuchengust) and [https://doi.org/10.5281/zenodo.832417](https://doi.org/10.5281/zenodo.832417).
Appendix A: Equations of regression techniques

Ordinary Least Square (OLS) calculation steps.

First calculate average of observed $X_i$ and $Y_i$.

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \quad (A1)$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} \quad (A2)$$

Then calculate $S_{xx}$ and $S_{yy}$.

$$S_{xx} = \sum_{i=1}^{N} (X_i - \bar{X})^2 \quad (A3)$$

$$S_{yy} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \quad (A4)$$

OLS slope and intercept can be obtained from,

$$k = \frac{S_{yy}}{S_{xx}} \quad (A6)$$

$$b = \bar{Y} - k\bar{X} \quad (A7)$$

Deming regression (DR) calculation steps (York, 1966).

Besides $S_{xx}$ and $S_{yy}$ as shown above, $S_{xy}$ can be calculated from,

$$S_{xy} = \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \quad (A8)$$

DR slope and intercept can be obtained from,

$$k = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \quad (A9)$$

$$b = \bar{Y} - k\bar{X} \quad (A10)$$

York regression (YR) iteration steps (York et al., 2004).

Slope by OLS can be used as the initial $k$ in $W_i$ calculation.

$$W_i = \frac{\omega(X_i)\omega(Y_i)}{\omega(X_i) + k^2 \omega(Y_i) - 2kr_i\sqrt{\omega(X_i)\omega(Y_i)}} \quad (A11)$$
Then calculate $\beta_i$.

$$
\beta_i = W_i \left[ \frac{U_i}{\omega(Y_i)} + \frac{k V_i}{\omega(X_i)} - \frac{r_i}{\sqrt{\omega(X_i) \omega(Y_i)}} \right] \quad (A14)
$$

Slope and intercept can be obtained from,

$$
k = \frac{\sum_{i=1}^{N} W_i \beta_i V_i}{\sum_{i=1}^{N} W_i \beta_i U_i} \quad (A15)
$$

$$
b = \bar{Y} - k \bar{X} \quad (A16)
$$

Since $W_i$ and $\beta_i$ are functions of $k$, $k$ must be solved iteratively by repeating A11 to A15. If the difference between the $k$ obtained from A15 and the $k$ used in A11 satisfies the predefined tolerance $(k_{i+1} - k_i < e^{-15})$, the calculation is considered as converged. The calculation is straightforward and usually converged in 10 iterations. For example, the iteration count on the data set of Chu (2005) is around 6.

**Data availability.** OC, EC and $\sigma_{abs}$ data used in this study are available from the corresponding authors upon request. The computer programs used for data analysis and visualization in this study are available in Wu (2017a–c).

**Competing interests.** The authors declare that they have no conflict of interest.

**Acknowledgements**

This work is supported by the National Natural Science Foundation of China (Grant No. 41605002, 41475004 and 21607056), NSFC of Guangdong Province (Grant No. 2015A030313339), Guangdong Province Public Interest Research and Capacity Building Special Fund (Grant No. 2014B020216005). The author would like to thank Dr. Bin Yu Kuang at HKUST for discussion on mathematics and Dr. Stephen M Griffith at HKUST for valuable comments.
References


Table 1. Summary of abbreviations and symbols.

<table>
<thead>
<tr>
<th>Abbreviation/symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>a dimensionless adjustable factor to control the position of $y_{one}$ curve on the concentration axis</td>
</tr>
<tr>
<td>$b$</td>
<td>intercept in linear regression</td>
</tr>
<tr>
<td>$\beta_i, U_i, V_i, W_i$</td>
<td>intermediates in York regression calculations</td>
</tr>
<tr>
<td>$y_{unc}$</td>
<td>fractional measurement uncertainties relative to the true concentration (%)</td>
</tr>
<tr>
<td>DR</td>
<td>Deming regression</td>
</tr>
<tr>
<td>$\epsilon_{EC}, \epsilon_{POC}$</td>
<td>absolute measurement uncertainties of EC and POC</td>
</tr>
<tr>
<td>EC</td>
<td>elemental carbon</td>
</tr>
<tr>
<td>$EC_{measured}$</td>
<td>numerically synthesized true EC concentration without measurement uncertainty</td>
</tr>
<tr>
<td>$EC_{true}$</td>
<td>numerically synthesized true EC concentration without measurement uncertainty</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$\omega(X_i)$ to $\omega(Y_i)$ ratio in Deming regression</td>
</tr>
<tr>
<td>$k$</td>
<td>slope in linear regression</td>
</tr>
<tr>
<td>LOD</td>
<td>limit of detection</td>
</tr>
<tr>
<td>MT</td>
<td>Mersenne twister pseudorandom number generator</td>
</tr>
<tr>
<td>OC</td>
<td>organic carbon</td>
</tr>
<tr>
<td>OC/EC</td>
<td>OC to EC ratio</td>
</tr>
<tr>
<td>$(OC/EC)_{pri}$</td>
<td>primary OC/EC ratio</td>
</tr>
<tr>
<td>OC$_{non-comb}$</td>
<td>OC from non-combustion sources</td>
</tr>
<tr>
<td>ODR</td>
<td>orthogonal distance regression</td>
</tr>
<tr>
<td>OLS</td>
<td>ordinary least squares regression</td>
</tr>
<tr>
<td>POC</td>
<td>primary organic carbon</td>
</tr>
<tr>
<td>$POC_{comb}$</td>
<td>numerically synthesized true POC from combustion sources (well correlated with $EC_{true}$), measurement uncertainty not considered</td>
</tr>
<tr>
<td>$POC_{non-comb}$</td>
<td>numerically synthesized true POC from non-combustion sources (independent of $EC_{true}$) without measurement uncertainty</td>
</tr>
<tr>
<td>$POC_{measured}$</td>
<td>POC with measurement error ($POC_{true} + \epsilon_{POC}$)</td>
</tr>
<tr>
<td>$\sigma_{x_i}, \sigma_{y_i}$</td>
<td>the standard deviation of the error in measurement of $X_i$ and $Y_i$</td>
</tr>
<tr>
<td>$r_{xy}$</td>
<td>correlation coefficient between errors in $X_i$ and $Y_i$ in YR</td>
</tr>
<tr>
<td>$S$</td>
<td>sum of squared residuals</td>
</tr>
<tr>
<td>SOC</td>
<td>secondary organic carbon</td>
</tr>
<tr>
<td>$\tau$</td>
<td>parameter in the sine function of Chu (2005) that adjusts the width of each peak</td>
</tr>
<tr>
<td>$\phi$</td>
<td>parameter in the sine function of Chu (2005) that adjusts the phase of the curve</td>
</tr>
<tr>
<td>WODR</td>
<td>weighted orthogonal distance regression</td>
</tr>
<tr>
<td>$\bar{X}, \bar{Y}$</td>
<td>average of $X_i$ and $Y_i$</td>
</tr>
<tr>
<td>YR</td>
<td>York regression</td>
</tr>
<tr>
<td>$\omega(X_i), \omega(Y_i)$</td>
<td>inverse of $\sigma_{x_i}$ and $\sigma_{y_i}$, used as weights in DR calculation.</td>
</tr>
</tbody>
</table>
Table 2. Summary of six regression approaches comparison with 5000 runs for 18 cases.

<table>
<thead>
<tr>
<th>Case</th>
<th>Data scheme</th>
<th>True Slope</th>
<th>True Intercept</th>
<th>R² (X, Y) Measurement error</th>
<th>OLS</th>
<th>DR</th>
<th>ODR</th>
<th>WODR</th>
<th>YR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Slope</td>
<td>Intercept</td>
<td>Slope</td>
<td>Intercept</td>
<td>Slope</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0.67±0.03</td>
<td>LOGDc=1, LOGDc=1</td>
<td>2.94±0.14</td>
<td>5.84±0.78</td>
<td>4.27±0.27</td>
<td>-1.45±1.36</td>
<td>4.01±0.25</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0.67±0.04</td>
<td>LOGDc=1, LOGDc=1</td>
<td>2.95±0.15</td>
<td>8.83±0.80</td>
<td>4.32±0.28</td>
<td>1.28±1.43</td>
<td>4.01±0.26</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>3</td>
<td>0.95±0.01</td>
<td>LOGDc=0.5, LOGDc=0.5</td>
<td>3.83±0.08</td>
<td>0.95±0.40</td>
<td>4.03±0.09</td>
<td>-0.18±0.44</td>
<td>4±0.09</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.78±0.02</td>
<td>YOc=30%</td>
<td>3.32±0.20</td>
<td>3.77±0.90</td>
<td>4.75±0.30</td>
<td>-4.14±1.36</td>
<td>4±0.25</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.69±0.04</td>
<td>YOc=30%</td>
<td>3.31±0.22</td>
<td>6.79±1.02</td>
<td>4.95±0.31</td>
<td>-2.26±1.48</td>
<td>3.99±0.26</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0.76±0.01</td>
<td></td>
<td>3.22±0.03</td>
<td>4.3±0.14</td>
<td>4.17±0.04</td>
<td>-0.94±0.18</td>
<td>4±0.03</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0.76±0.01</td>
<td></td>
<td>3.22±0.03</td>
<td>7.29±0.14</td>
<td>4.2±0.04</td>
<td>1.88±1.18</td>
<td>4±0.03</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0.75±0.01</td>
<td></td>
<td>0.41±0.00</td>
<td>0.36±0.02</td>
<td>0.46±0.01</td>
<td>0.23±0.03</td>
<td>0.5±0.01</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0.76±0.01</td>
<td></td>
<td>0.43±0.01</td>
<td>3.36±0.03</td>
<td>0.5±0.01</td>
<td>3.02±0.04</td>
<td>0.49±0.01</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.56±0.01</td>
<td></td>
<td>0.87±0.01</td>
<td>0.72±0.05</td>
<td>1±0.01</td>
<td>0±0.06</td>
<td>1±0.01</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0.76±0.01</td>
<td></td>
<td>0.87±0.01</td>
<td>3.72±0.05</td>
<td>1.09±0.01</td>
<td>2.52±0.07</td>
<td>0.99±0.01</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0.66±0.01</td>
<td></td>
<td>3.48±0.04</td>
<td>2.87±0.18</td>
<td>4.53±0.05</td>
<td>-2.94±0.24</td>
<td>4±0.05</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0.76±0.01</td>
<td></td>
<td>3.48±0.04</td>
<td>5.87±0.19</td>
<td>4.67±0.05</td>
<td>-0.67±0.26</td>
<td>3.98±0.05</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0.73±0.01</td>
<td></td>
<td>0.44±0.01</td>
<td>0.55±0.03</td>
<td>0.45±0.01</td>
<td>0.26±0.03</td>
<td>0.5±0.01</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0.54±0.01</td>
<td></td>
<td>0.44±0.01</td>
<td>3.54±0.04</td>
<td>0.5±0.01</td>
<td>2.98±0.04</td>
<td>0.5±0.01</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0.54±0.01</td>
<td></td>
<td>0.8±0.01</td>
<td>1.07±0.04</td>
<td>1±0.01</td>
<td>0±0.05</td>
<td>1±0.01</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>1</td>
<td>0.59±0.01</td>
<td></td>
<td>0.8±0.01</td>
<td>4.07±0.05</td>
<td>1.07±0.01</td>
<td>2.62±0.07</td>
<td>1±0.01</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>3</td>
<td>0.59±0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 1.** (a) Example $\gamma_{\text{unc-linear}}$ curves by different $\alpha$ values (Eq. (17)). The X axis is concentration (normalized by LOD) in log scale and Y axis is $\gamma_{\text{unc}}$. Black, blue and green line represent $\alpha$ equal to 1, 0.5 and 0.3, respectively, corresponding to the $\gamma_{\text{unc-linear}}$ at LOD level equals to 100%, 50% and 30%, respectively. The red line represents $\gamma_{\text{unc-linear}}$ of 10%. (b) Example of measurement uncertainty generation of $\gamma_{\text{unc-linear}}$ and $\gamma_{\text{unc-nonlinear}}$. The blue circles represent $\gamma_{\text{unc-nonlinear}}$ following Eq. (17) ($LOD_{EC} = 1$, $a_{EC} = 1$). The red circles represent $\gamma_{\text{unc-linear}}$ (30%).
Data generation steps by MT

1) Generate $E_{C_{\text{true}}}$ by MT

2) Compute $POC_{\text{comb}}$

3) Compute $POC_{\text{true}}$

(4-a) Compute $\varepsilon_{POC}$ and $\varepsilon_{EC}$ with $\gamma_{unc\text{-nonlinear}}$

$$-\frac{\log_{10}EC_{true} \times a_{POC} \times POC_{true}}{POC_{comb}} \leq \varepsilon_{POC} \leq +\frac{\log_{10}EC_{true} \times a_{POC} \times POC_{true}}{POC_{comb}}$$


(4-b) Compute $\varepsilon_{POC}$ and $\varepsilon_{EC}$ with $\gamma_{unc\text{-linear}}$

$$-\gamma_{POC_{comb}}POC_{true} \leq \varepsilon_{POC} \leq +\gamma_{POC_{comb}}POC_{true}$$

$$-\gamma_{EC_{true}}EC_{true} \leq \varepsilon_{EC} \leq +\gamma_{EC_{true}}EC_{true}$$

5) Compute $POC_{\text{measured}}$ and $EC_{\text{measured}}$

Figure 2. Flowchart of data generation steps using MT.
Figure 3. POC_{comb} and EC_{true} data generated by the sine functions of Chu (2005). (a) Time series of the 120 data points for POC_{comb} and EC_{true}. (b) Scatter plot of POC_{comb} vs. EC_{true}.

\[ POC_{comb} = 14 + 12(\sin \left( \frac{x}{7} \right) + \sin(x - \phi)) \]

\[ EC_{true} = 3.5 + 3(\sin \left( \frac{x}{7} \right) + \sin(x - \phi)) \]
**Comparison study design**

**Figure 4.** Overview of the comparison study design.

- Data generation using sine function
- Data generation using MT
- Nonlinear $γ_{Unc}$
- Linear $γ_{Unc}$
- Six regression approaches
- Nonlinear $γ_{Unc}$
- Linear $γ_{Unc}$
- Six regression approaches

- OLS
- DR ($λ = 1$)
- DR ($λ = \frac{w(x_i)}{w(\bar{y})}$)
- ODR
- WODR
- YR
Figure 5. Regression results on synthetic data, case 1 (Slope=4, Intercept=0, $LOD_{POC}=1, LOD_{EC}=1, a_{POC}=1, a_{EC}=1$, $R^2$ (POC, EC) =0.67±0.03). The scatter plots demonstrate regression examples from a single run. The box plots show the distribution of regressed slopes and intercepts from 5000 runs of six regression approaches. The dashed line in orange and peachblow represent true slope and intercept, respectively.
Figure 6. Slope and intercept biases by different regression schemes in two test scenarios (A and B) in which the assumed error for one of the regression variables deviates from the actual measurement error. In Test A data generation, \( y_{\text{unc,x}} \) is fixed at 30% and \( y_{\text{unc,y}} \) is varied between 1 ~ 50%. In Test B, \( y_{\text{unc,x}} \) varies between 1 ~ 50% and \( y_{\text{unc,y}} \) is fixed at 30%. The "true" measurement error for regression is 10% for both X and Y. (a) Slopes biases as a function of \( y_{\text{unc,y}} \) in Test A. (b) Intercepts biases as a function of \( y_{\text{unc,y}} \) in Test A. (c) Slopes biases as a function of \( y_{\text{unc,x}} \) in Test B. (d) Intercepts biases as a function of \( y_{\text{unc,x}} \) in Test B.
Figure 7. Regression results using ambient $\sigma_{abs520}$ and EC data from a suburban site in Guangzhou, China.
Figure 8. The user interface of Scatter Plot Igor program. The program and its operation manual are available from: https://doi.org/10.5281/zenodo.832417.