Supplement of Atmos. Meas. Tech., 11, 1233–1250, 2018
https://doi.org/10.5194/amt-11-1233-2018-supplement

Atmospheric Measurement Techniques
Open Access

Supplement of

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

Cheng Wu and Jian Zhen Yu

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

This document contains three supporting tables, nine supporting figures.

1 Comparison of three York regression implementations

A variety of York regression implementations are compared using the Pearson's data with York's weights according to York (1966) (abbreviated as "PY data" hereafter). The dataset is given in Table S2.Three York regression implementations are compared using the PY data, including spreadsheet by Cantrell (2008), Igor program by this study and a commercial software (OriginPro™ 2017). The three York regression implementations yield identical slope and intercept as shown in the highlighted areas (in red) in Figure S6. These crosscheck results suggest that the codes in our Igor program can retrieve consistent slopes and intercepts as other proven programs did.

2 Impact of two primary sources in OC/EC regression

A sampling site is often influenced by multiple combustion sources in the real atmosphere. In section 1 and 2 of the main text we evaluate the performance of OLS, DR, WODR and YR in scenarios of two primary sources and arbitrarily dictate that the $(OC/EC)_{pri}$ of source 1 is lower than that of source 2. By varying f_{EC1} (proportion of source 1 EC to total EC) from test to test, the effect of different mixing ratios of the two sources can be examined. Two scenarios are considered (Wu and Yu, 2016): two correlated primary sources and two independent primary sources. Common configurations include: EC_{total}=2 μgC m^{-3}; f_{EC1} varies from 0 to 100%; ratio of the two OC/EC_{pri} values ($\gamma_{\_pri}$) vary in the range of 2~8. Studies by Chu (2005) and Saylor et al. (2006) both suggest ratio of averages (ROA) being the best estimator of the expected primary OC/EC ratio when SOC is zeroed. Since the overall OC/EC_{pri} from the two sources varies by $\gamma_{\_pri}$, ROA is considered as the reference OC/EC_{pri} to be compared with slope regressed by of OLS, DR, WODR and YR. The abbreviations used for the two primary sources study are listed in Table S3.

2.1 Impact of two correlated primary sources

Simulations considering two correlated primary sources are performed, to examine the effect on bias in the regression methods. The basic configuration is: $(OC/EC)_{pri1}$=0.5, $(OC/EC)_{pri2}$=5, γ_{Unc}=30%, N=8000, intercept=0, and the following terms are compared:

ratio of averages (ROA here refers to the ratio of averaged OC to averaged EC, which is considered as the true value of slope when intercept=0), DR, WODR, WODR' (through origin) and OLS. As shown in Figure S7, when R^2 (EC1 vs. EC2) is very high, DR, WODR and WODR' can provide a result consistent with ROA. If the R^2 decreases, the bias of the slope and intercept in DR and WODR is larger. OLS constantly underestimates the slope.

2.2 Impact of two independent primary sources

Simulations of two independent primary sources are also conducted. If RSD_{EC1}=RSD_{EC2}, slopes and intercepts may be either overestimated or underestimated (Figure S8), and the degree of bias depends on the magnitude of RSD_{EC1} and RSD_{EC2}. Larger RSD results in larger bias. Uneven RSD between two sources leads to even more bias (Figure S8 a and b). The degree of bias also shows dependence on γ_pri. If γ_pri decreases, the bias becomes smaller (FigureS8 c~f). These results indicate that the scenario with two independent primary sources poses a challenge to $(OC/EC)_{pri}$ estimation by linear regression.

For the EC tracer method, if EC comes from two primary sources and contribution of the two sources is comparable, the regression slope is no longer suitable for $(OC/EC)_{pri}$ estimation and the subsequent SOC calculation, and making EC a mixture that violates the property of a tracer. For such a situation, pre-separation of EC into individual sources by other tracers (if available) by the Minimum R Squared (MRS) method can provide unbiased SOC estimation results (Wu and Yu, 2016).

3 Igor programs for error in variables linear regression and simulated OC EC data generation using MT

An Igor Pro (WaveMetrics, Inc. Lake Oswego, OR, USA) based program (Scatter plot) with graphical user interface (GUI) is developed to make the linear regression feasible and user friendly (Figure 8). The program includes Deming and York algorithm for linear regression, which considers uncertainties in both X and Y, that is more realistic for atmospheric applications. 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.

Another program using MT can generate simulated OC and EC concentration through user defined parameters via GUI as shown in Figure S9.

Both Igor programs and their operation manuals can be downloaded from the following links:

https://sites.google.com/site/wuchengust

https://doi.org/10.5281/zenodo.832417

References

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, 5477-5487, 10.5194/acp-8-5477-2008, 2008.

Chu, S. H.: Stable estimate of primary OC/EC ratios in the EC tracer method, Atmos. Environ., 39, 1383-1392, 10.1016/j.atmosenv.2004.11.038, 2005.

Saylor, R. D., Edgerton, E. S., and Hartsell, B. E.: Linear regression techniques for use in the EC tracer method of secondary organic aerosol estimation, Atmos. Environ., 40, 7546-7556, 10.1016/j.atmosenv.2006.07.018, 2006.

Wu, C. and Yu, J. Z.: Determination of primary combustion source organic carbon-to-elemental carbon (OC/EC) ratio using ambient OC and EC measurements: secondary OC-EC correlation minimization method, Atmos. Chem. Phys., 16, 5453-5465, 10.5194/acp-16-5453-2016, 2016.

York, D.: Least-squares fitting of a straight line, Can. J. Phys., 44, 1079-1086, 10.1139/p66-090, 1966.

Table S1. Summary of the five linear regression techniques.

Approach	Sum of squared residuals (SSR)	Calculation
Ordinary least squares (OLS)	$\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{y_i} - \boldsymbol{Y_i})^2$	closed form
Orthogonal distance regression (ODR)	$\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} [(\boldsymbol{x_i} - \boldsymbol{X_i})^2 + (\boldsymbol{y_i} - \boldsymbol{Y_i})^2]$	iteration
Weighted orthogonal distance regression (WODR)	$\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} [(\boldsymbol{x_i} - \boldsymbol{X_i})^2 + (\boldsymbol{y_i} - \boldsymbol{Y_i})^2/\boldsymbol{\eta}]$	iteration
Deming regression (DR)	$\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} [\boldsymbol{\omega}(\boldsymbol{X_i})(\boldsymbol{x_i} - \boldsymbol{X_i})^2 + \boldsymbol{\omega}(\boldsymbol{Y_i})(\boldsymbol{y_i} - \boldsymbol{Y_i})^2]$	closed form
York regression (YR)	$\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \left[\boldsymbol{\omega}(\boldsymbol{X_i})(\boldsymbol{x_i} - \boldsymbol{X_i})^2 - \boldsymbol{2r_i}\sqrt{\boldsymbol{\omega}(\boldsymbol{X_i})\boldsymbol{\omega}(\boldsymbol{Y_i})}(\boldsymbol{x_i} - \boldsymbol{X_i})(\boldsymbol{y_i} - \boldsymbol{Y_i}) + \boldsymbol{\omega}(\boldsymbol{Y_i})(\boldsymbol{y_i} - \boldsymbol{Y_i})^2\right] \frac{\boldsymbol{1}}{\boldsymbol{1} - \boldsymbol{r_i}^2}$	iteration

Table S2. Pearson's data with York's weights according to York (1966).

X_i	$\omega(X_i)$	Y_i	$\omega(Y_i)$
0	1000	5.9	1
0.9	1000	5.4	1.8
1.8	500	4.4	4
2.6	800	4.6	8
3.3	200	3.5	20
4.4	80	3.7	20
5.2	60	2.8	70
6.1	20	2.8	70
6.5	1.8	2.4	100
7.4	1	1.5	500

Table S3. Abbreviations used in two primary sources study.

Abbreviation	Definition
EC_1,EC_2	EC from source 1 and source 2 in the two sources scenario
f_{EC1}	fraction of EC from source 1 to the total EC
ROA	ratio of averages (Y to X, e.g., averaged OC to averaged EC)
γ_pri	ratio of the $(OC/EC)_{pri}$ of source 2 to source 1
RSD	relative standard deviation
RSD_{EC}	RSD of EC
ε_{EC} , ε_{OC}	measurement uncertainty of EC and OC
γ_{unc}	relative measurement uncertainty
$\gamma_{\_RSD}$	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

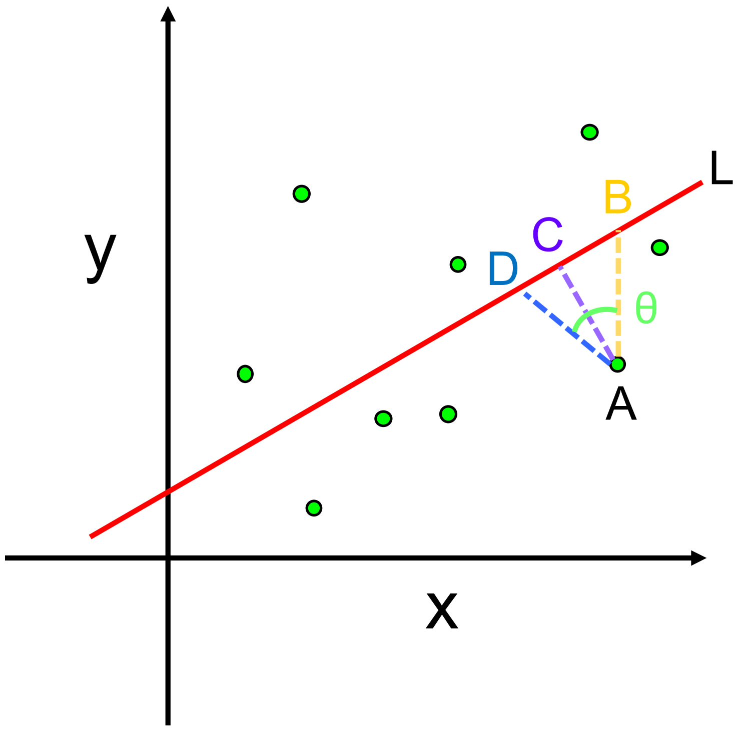


Figure S1. Relationships between data point A and fitting line L. Fitting line by OLS minimizes the distance of AB (AB is perpendicular to the x axis). Fitting line by ODR and DR ($\lambda = 1$) minimizes the distance of AC (AC is perpendicular to L). Fitting line by WODR, DR ($\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$) and YR minimizes the distance of AD. AD has a θ degree angle relative to AB and the θ depends on the weights of measurement errors in Y and X.

Data generation steps by the sine functions of Chu (2005)

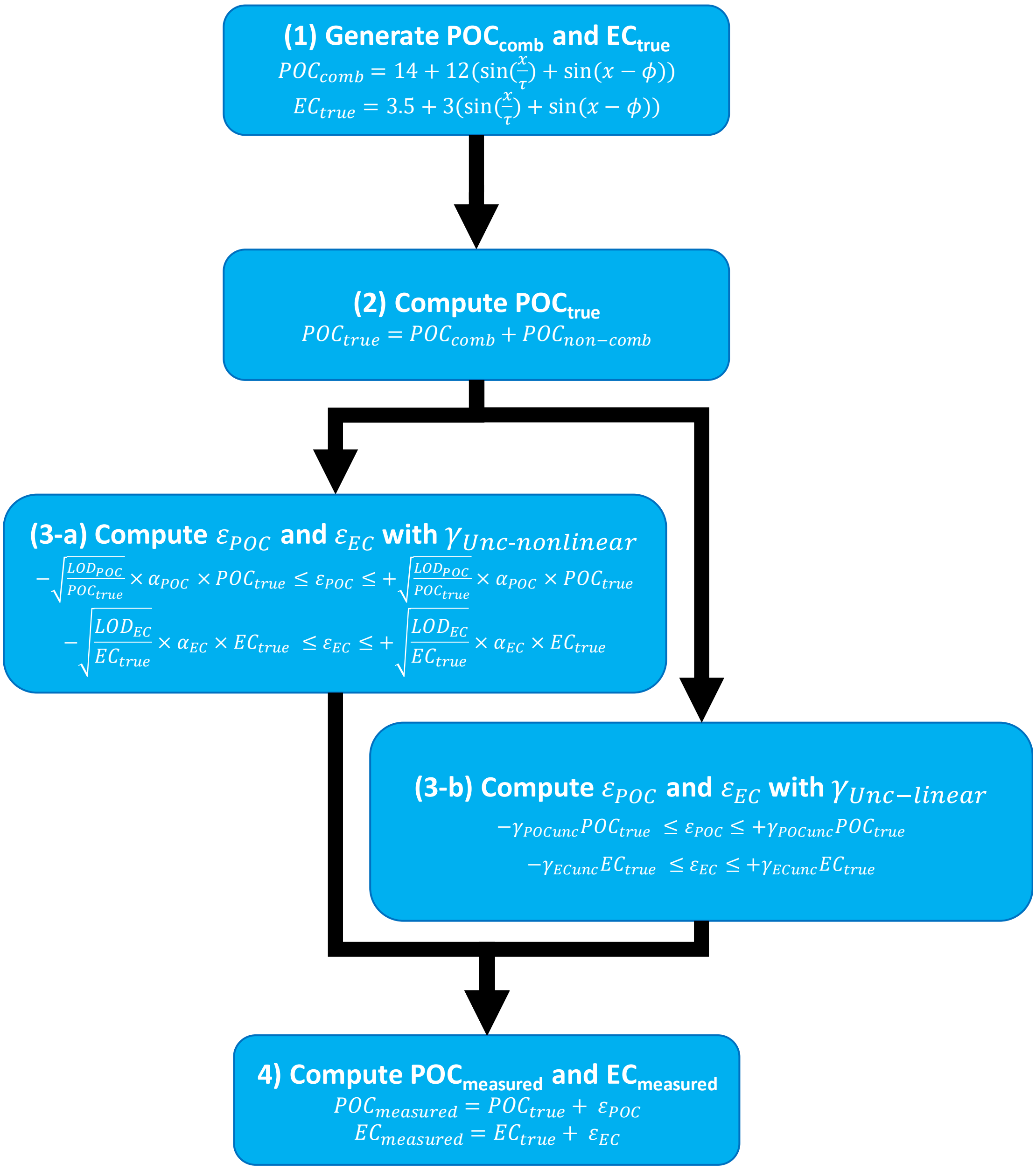


Figure S2. Flowchart of data generation steps using the sine functions of Chu (2005).

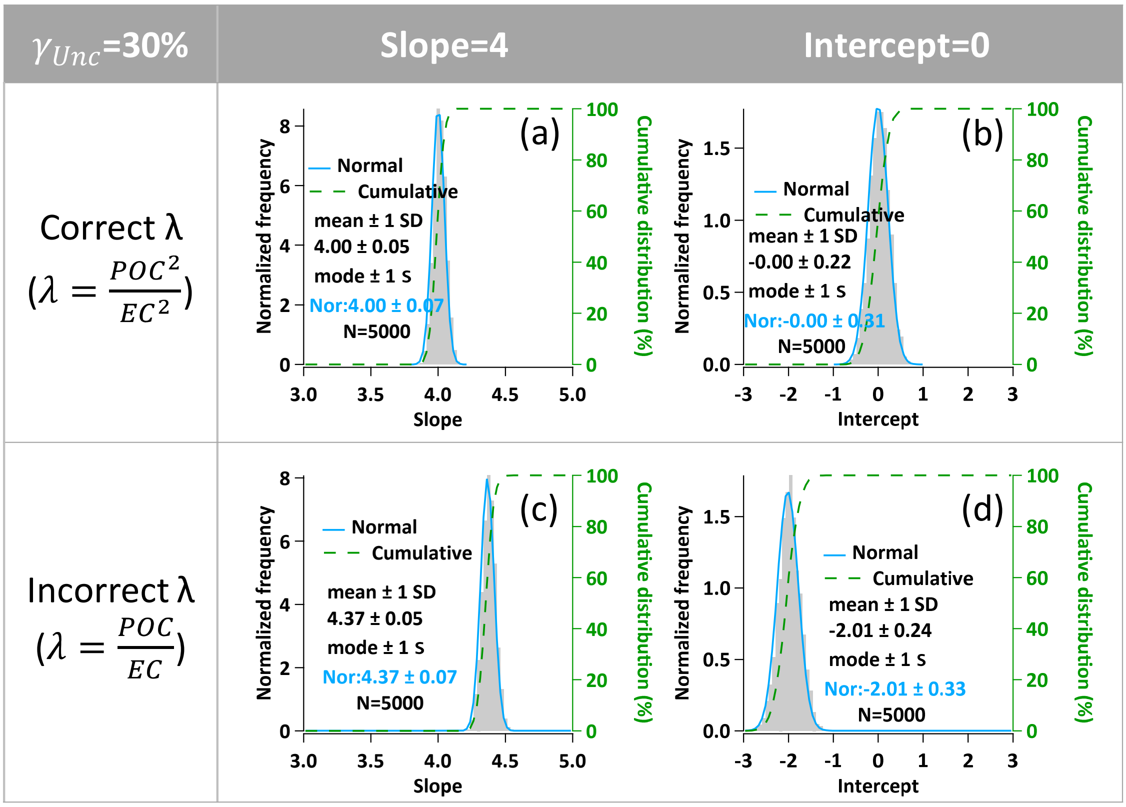


Figure S3. Example of bias in slope and intercept due to improper λ assignment. Data generation: Slope=4, Intercept=0; linear γ_{Unc} (30%). (a)&(b) Slopes and intercepts when proper λ is input following linear γ_{Unc} $(\lambda = \frac{POC^2}{EC^2})$; (c)&(d) Slopes and intercepts when improper λ is input following non-linear γ_{Unc} $(\lambda = \frac{POC}{EC})$.

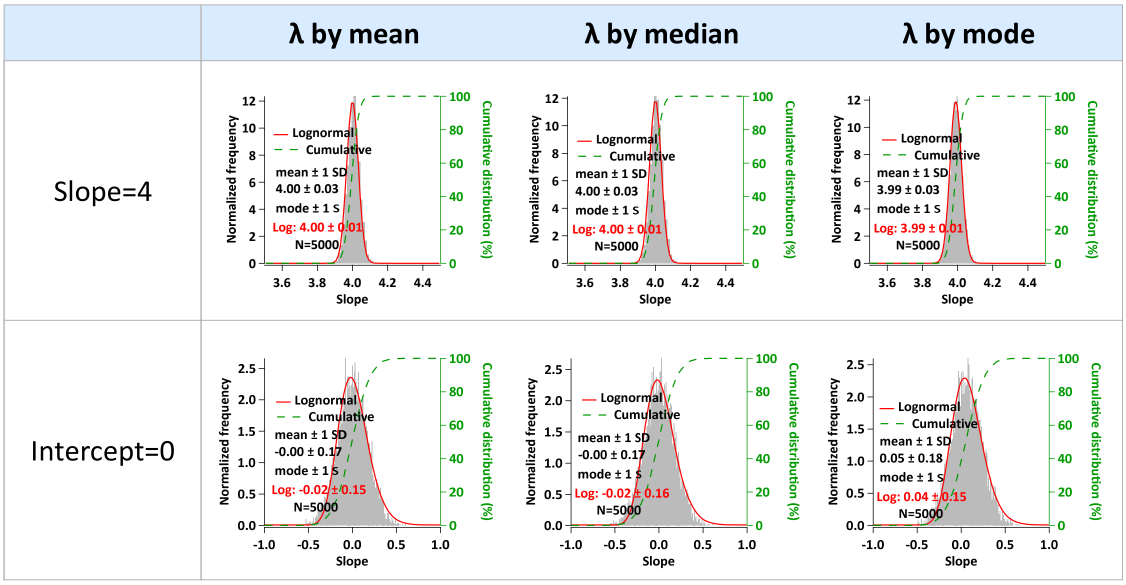


Figure S4. Sensitivity tests of λ calculated by mean, median and mode.

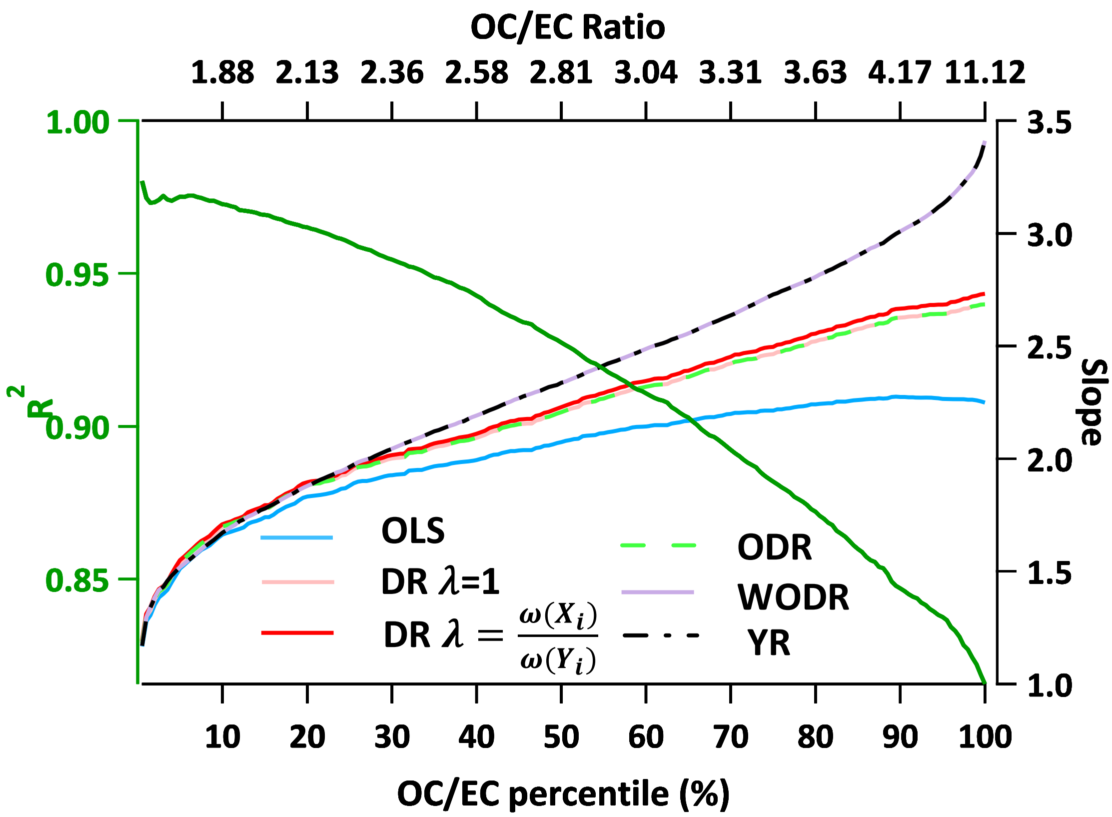


Figure S5. Regression slopes as a function of OC/EC percentile. OC/EC percentile range from 0.5% to 100%, with an interval of 0.5%.

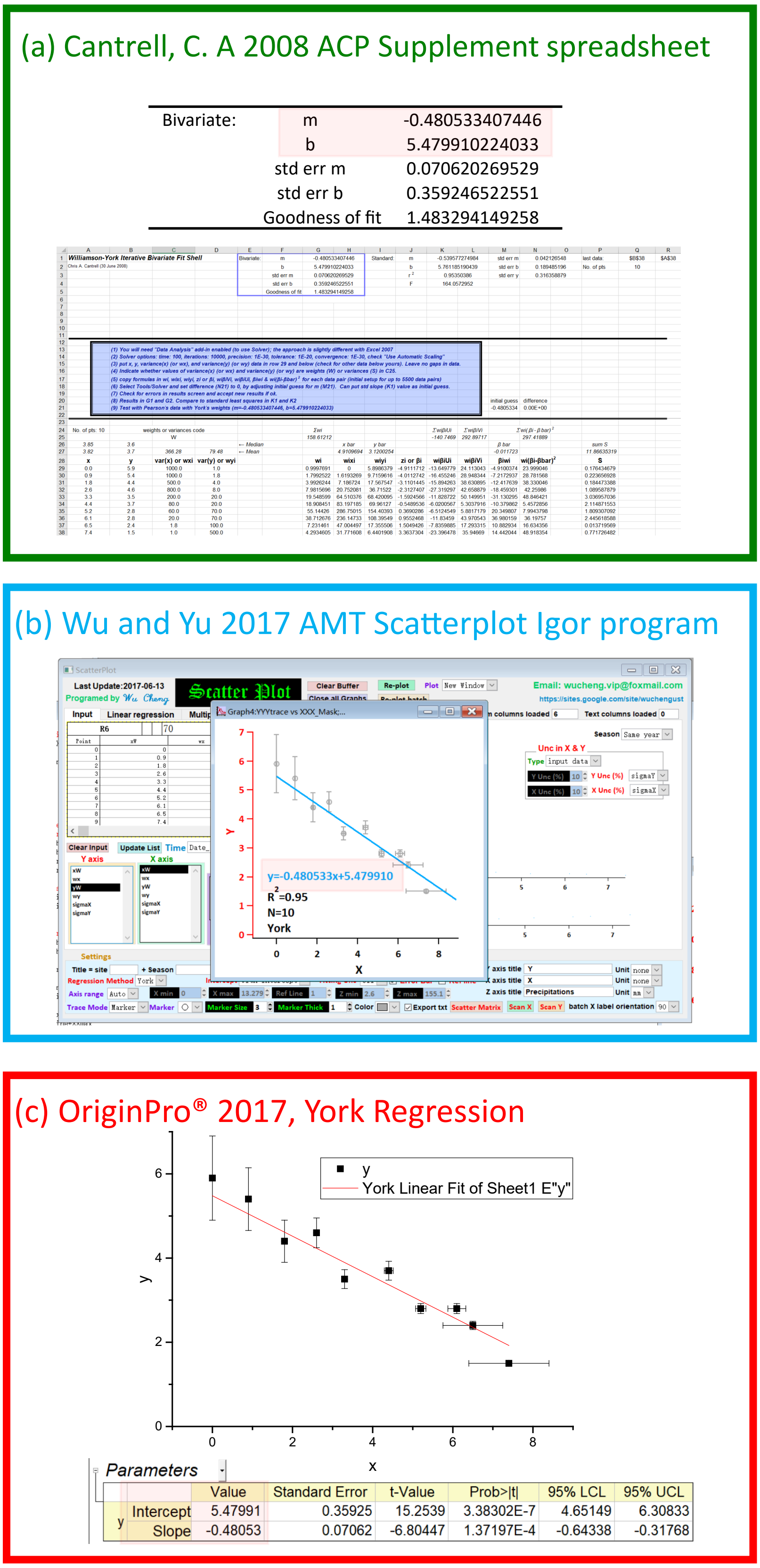


Figure S6. York regression implementations comparison using data shown in Table S2, including (a) spreadsheet by Cantrell (2008), (b) Igor program by this study and (c) a commercial software (OriginPro® 2017).

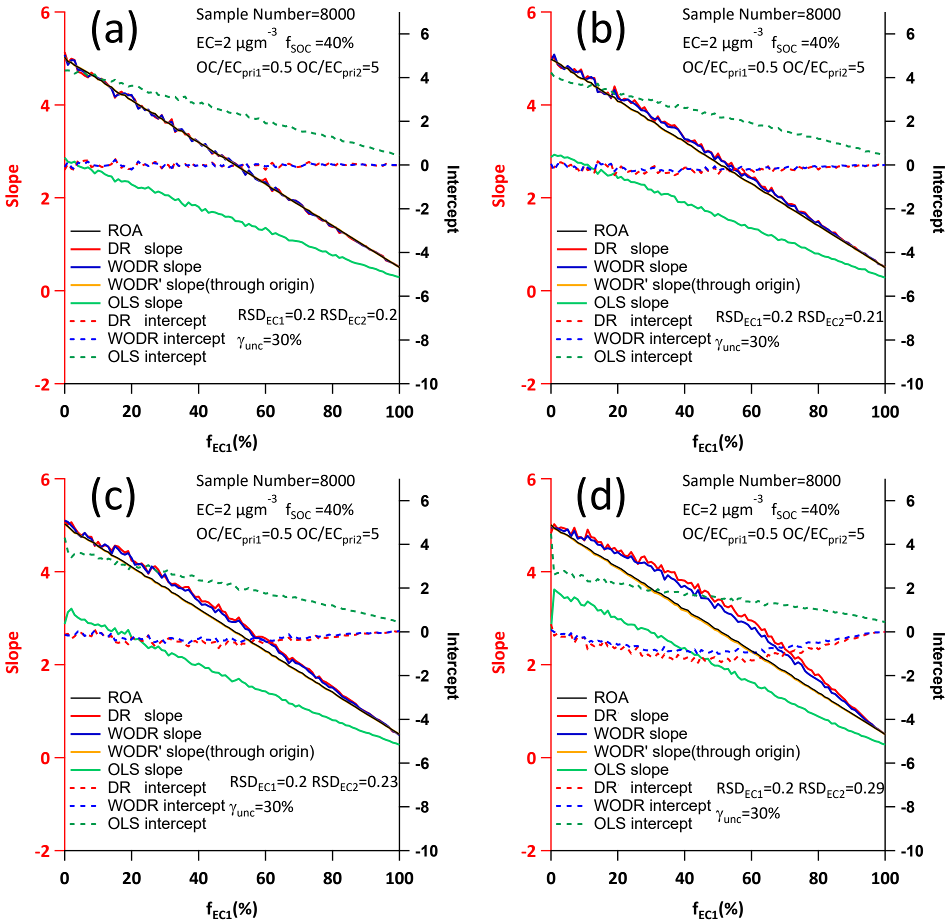


Figure S7. Study of two correlated sources scenario by different R^2 between the two sources. (a) $R^2 = 1$ (b) $R^2 = 0.86$ (c) $R^2 = 0.75$ (d) $R^2 = 0.49$.

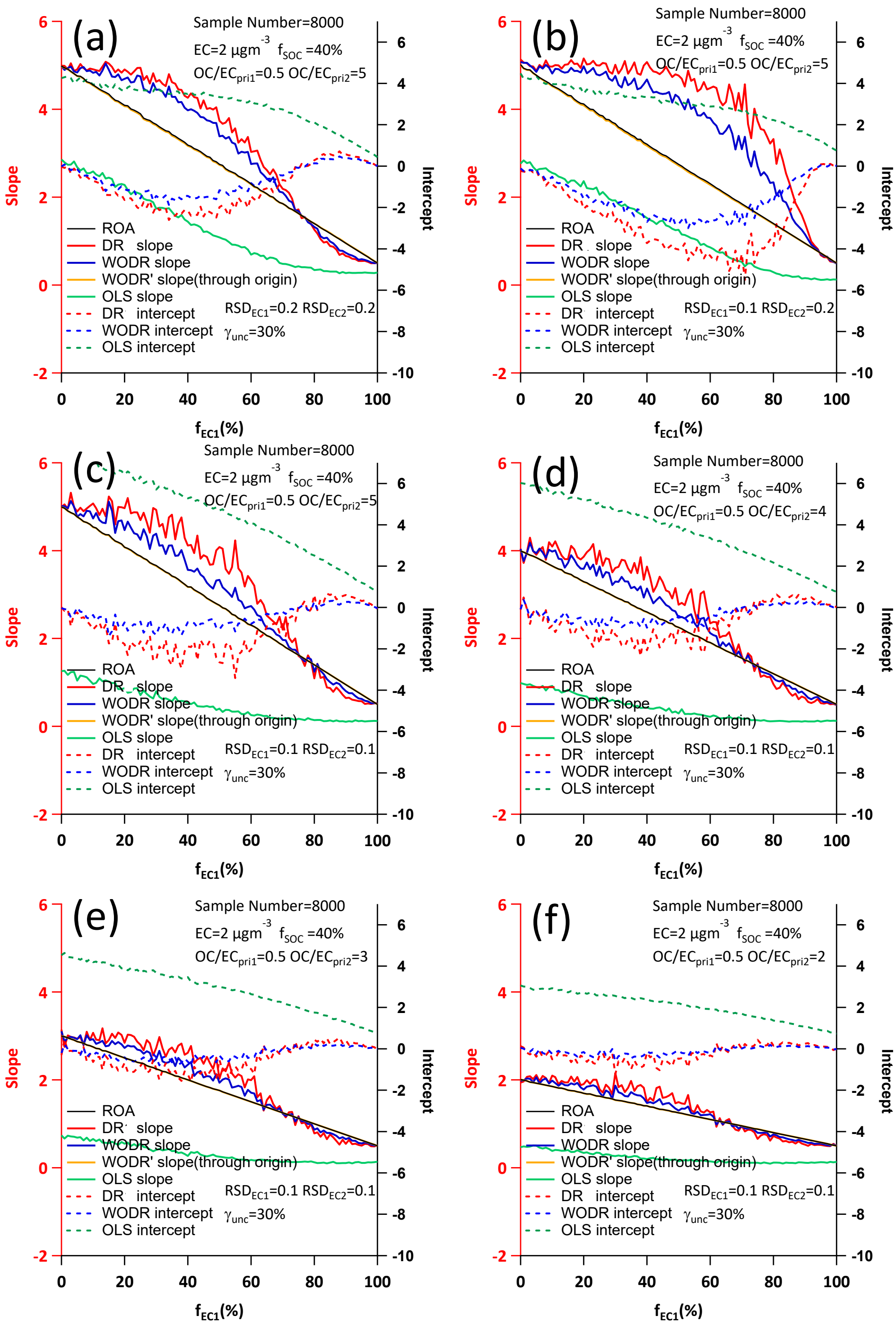
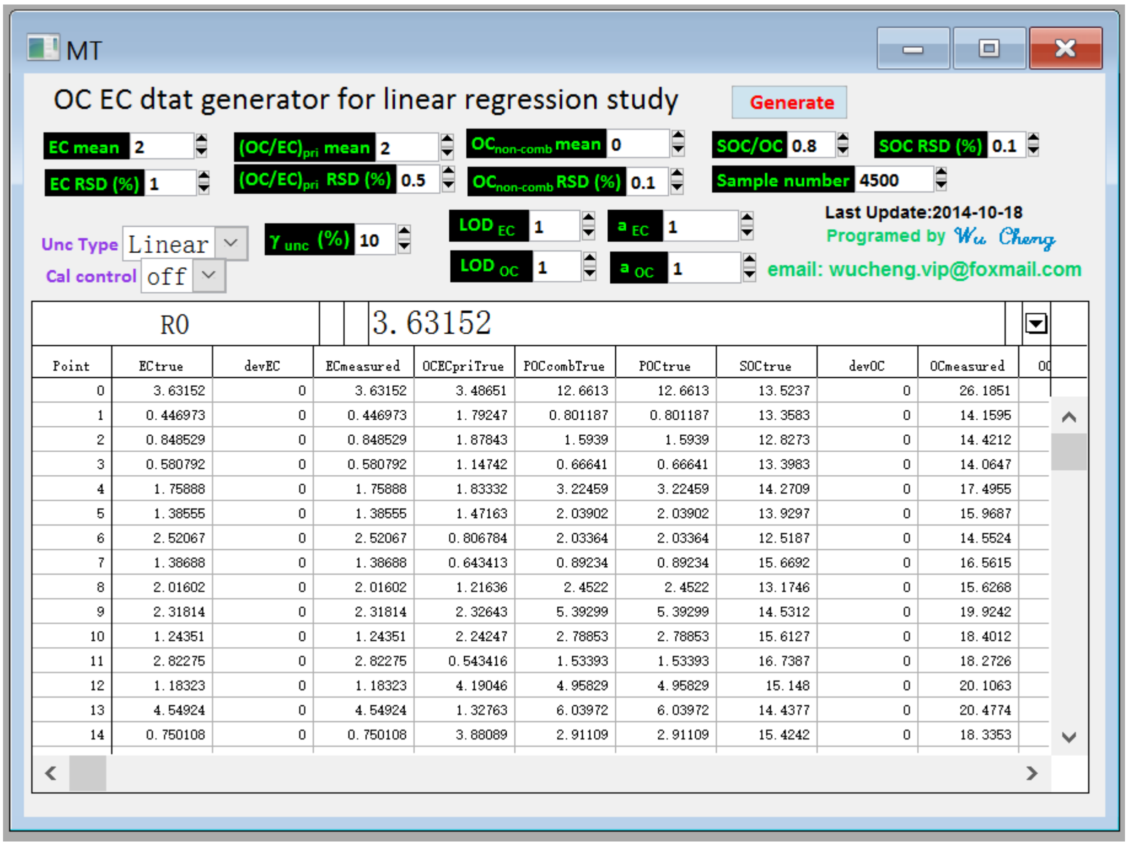


Figure S8. Study of two independent sources scenario by different parameters. (a) $\gamma\_pri=10$, $RSD_{EC1}=0.2$, $RSD_{EC2}=0.2$ (b) $\gamma\_pri=10$, $RSD_{EC1}=0.1$, $RSD_{EC2}=0.2$ (c) $\gamma\_pri=10$, $RSD_{EC1}=0.1$, $RSD_{EC2}=0.1$ (d) $\gamma\_pri=8$, $RSD_{EC1}=0.1$, $RSD_{EC2}=0.1$(e) $\gamma\_pri=6$, $RSD_{EC1}=0.1$, $RSD_{EC2}=0.1$ (f) $\gamma\_pri=4$, $RSD_{EC1}=0.1$, $RSD_{EC2}=0.1$.



Point	ECtrue	devEC	ECmeasured	OCECpriTrue	POCcombTrue	POCtrue	SOCtrue	devOC	OCmeasured
0	3. 63152	0	3. 63152	3. 48651	12. 6613	12. 6613	13. 5237	0	26. 1851
1	0. 446973	0	0. 446973	1. 79247	0. 801187	0. 801187	13. 3583	0	14. 1595
2	0. 848529	0	0. 848529	1. 87843	1. 5939	1. 5939	12. 8273	0	14. 4212
3	0. 580792	0	0. 580792	1. 14742	0. 66641	0. 66641	13. 3983	0	14. 0647
4	1. 75888	0	1. 75888	1. 83332	3. 22459	3. 22459	14. 2709	0	17. 4955
5	1. 38555	0	1. 38555	1. 47163	2. 03902	2. 03902	13. 9297	0	15. 9687
6	2. 52067	0	2. 52067	0. 806784	2. 03364	2. 03364	12. 5187	0	14. 5524
7	1. 38688	0	1. 38688	0. 643413	0. 89234	0. 89234	15. 6692	0	16. 5615
8	2. 01602	0	2. 01602	1. 21636	2. 4522	2. 4522	13. 1746	0	15. 6268
9	2. 31814	0	2. 31814	2. 32643	5. 39299	5. 39299	14. 5312	0	19. 9242
10	1. 24351	0	1. 24351	2. 24247	2. 78853	2. 78853	15. 6127	0	18. 4012
11	2. 82275	0	2. 82275	0. 543416	1. 53393	1. 53393	16. 7387	0	18. 2726
12	1. 18323	0	1. 18323	4. 19046	4. 95829	4. 95829	15. 148	0	20. 1063
13	4. 54924	0	4. 54924	1. 32763	6. 03972	6. 03972	14. 4377	0	20. 4774
14	0. 750108	0	0. 750108	3. 88089	2. 91109	2. 91109	15. 4242	0	18. 3353

Figure S9. MT Igor program. OC and EC data following log-normal distribution can be generated for statistical study purpose (no time series information). User can define mean and RSD of EC, $(OC/EC)_{pri}$, SOC/OC ratio, measurement uncertainty, sample size, etc. MT Igor program can be downloaded from the following link: https://sites.google.com/site/wuchengust.