Supplement of

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

Cheng Wu[1,2] and Jian Zhen Yu[3,4,5]

[1]Institute of Mass Spectrometer and Atmospheric Environment, Jinan University, Guangzhou 510632, China

[2]Guangdong Provincial Engineering Research Center for on-line source apportionment system of air pollution, Guangzhou 510632, China

[3]Division of Environment, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[4]Atmospheric Research Centre, Fok Ying Tung Graduate School, Hong Kong University of Science and Technology, Nansha, China

[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)

This document contains eight supporting tables, seven supporting figures, and discussion on the impact of two primary sources.

A sampling site is often impated by multiple combustion sources in the real atmosphere. In section 1 and 2 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 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 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 source varis 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 two primary sources study are listted in Table S8.

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 average (ROA, which is considered as the true value of slope when intercept=0), DR, WODR, WODR' (through origin) and OLS. As shown in Figure S5, 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 underestimate the slope.

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 S6), 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 S6 a&b). The degree of bias also shows dependence on γ_pri. If γ_pri decreases, the bias becomes smaller (FigureS6 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 7). The program includes Deming and York algorithm for linear regression, which consider uncertainties in both X and Y, that is more realistic for atmospheric applications. It packed with lots of useful features for data analysis and graph 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 S7.

Both Igor programs and their operation manuals can be downloaded from the following link: https://sites.google.com/site/wuchengust.

References

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.

Table S1. Comparison of regression techniques using the data generation scheme of Chu (2005) with 5000 runs considering different LOD and α configurations. Non-linear relative measurement uncertainty follows Eqs. (18) & (19). Unbiased (difference <5% to true value) slopes and intercepts are highlighted in grey.

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 3	OLS	3.83	±0.08	0.95	±0.40
Slope=4, Intercept=0	DR λ=1	4.03	±0.09	-0.18	±0.44
LOD_{POC}=0.5, LOD_{EC}=0.5 α_{POC}=0.5, α_{EC}=0.5	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.00	±0.09	0.00	±0.44
	ODR	4.03	±0.09	-0.18	±0.44
R^2 (POC,EC) = 0.95±0.01	WODR	4.00	±0.08	0.12	±0.37
	YR	4.00	±0.08	0.12	±0.37
Case 4	OLS	3.39	±0.15	3.34	±0.75
Slope=4, Intercept=0	DR1 λ=1	4.30	±0.21	-1.66	±1.06
LOD_{POC}=1, LOD_{EC}=0.5 α_{POC}=1, α_{EC}=1	DR1 $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.00	±0.19	-0.03	±0.99
	ODR	4.30	±0.21	-1.66	±1.06
R^2 (POC,EC) = 0.78±0.02	WODR	4.00	±0.17	0.33	±0.81
	YR	4.00	±0.17	0.33	±0.81

Table S2. Comparison of regression techniques using the data generation scheme of Chu (2005) with 5000 runs with linear measurement uncertainty following Eqs. (23) & (24). Unbiased (difference <5% to true value) slopes and intercepts are highlighted in grey.

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 5 Slope=4, Intercept=0 γ_{Unc}=30% R^2 (POC,EC) = 0.69±0.04	OLS	3.32	±0.20	3.77	±0.90
	DR λ=1	4.75	±0.30	-4.14	±1.36
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.01	±0.25	-0.04	±1.13
	ODR	4.75	±0.30	-4.14	±1.36
	WODR	4.00	±0.18	-0.01	±0.59
	YR	4.00	±0.18	-0.01	±0.59
Case 6 Slope=4, Intercept=3 γ_{Unc}=30% R^2 (POC,EC) = 0.66±0.04	OLS	3.31	±0.22	6.79	±1.02
	DR1 λ=1	4.95	±0.31	-2.26	±1.48
	DR1 $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	3.99	±0.26	3.05	±1.22
	ODR	4.95	±0.31	-2.26	±1.48
	WODR	4.01	±0.20	2.72	±0.74
	YR	4.01	±0.20	2.72	±0.74

Table S3. Comparison of regression techniques using MT data generation scheme with 5000 runs and a nonlinear measurement uncertainty following Eqs. (23) & (24).

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 9	OLS	0.43	±0.00	0.36	±0.02
Slope=0.5, Intercept=0	DR λ=1	0.46	±0.01	0.23	±0.03
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1.	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	0.50	±0.01	-0.00	±0.03
R^2 (POC,EC) = 0.76±0.01	ODR	0.46	±0.01	0.23	±0.03
	WODR	0.50	±0.00	-0.00	±0.01
	YR	0.50	±0.00	-0.00	±0.01
Case 10	OLS	0.43	±0.01	3.36	±0.03
Slope=0.5, Intercept=3	DR λ=1	0.50	±0.01	3.02	±0.04
LOD_{POC}=1, LOD_{EC}=1 a_{POC}=1, a_{EC}=1.	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	0.49	±0.01	3.05	±0.04
R^2 (POC,EC) = 0.56±0.01	ODR	0.50	±0.01	3.02	±0.04
	WODR	0.51	±0.01	2.73	±0.03
	YR	0.51	±0.01	2.73	±0.03

Table S4. Comparison of regression techniques using MT data generation scheme with 5000 runs, and a nonlinear measurement uncertainty following Eqs. (23) & (24).

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 11	OLS	0.87	±0.01	0.72	±0.05
Slope=1, Intercept=0	DR λ=1	1.00	±0.01	-0.00	±0.06
LOD_{POC}=1, LOD_{EC}=1	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	1.00	±0.01	-0.00	±0.06
a_{POC}=1, a_{EC}=1.	ODR	1.00	±0.01	-0.00	±0.06
R^2 (POC,EC) = 0.76±0.01	WODR	1.00	±0.01	-0.00	±0.02
	YR	1.00	±0.01	-0.00	±0.02
Case 12	OLS	0.87	±0.01	3.72	±0.05
Slope=1, Intercept=3	DR λ=1	1.09	±0.01	2.52	±0.07
LOD_{POC}=1, LOD_{EC}=1	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	0.99	±0.01	3.07	±0.06
a_{POC}=1, a_{EC}=1.	ODR	1.09	±0.01	2.52	±0.07
R^2 (POC,EC) = 0.66±0.01	WODR	1.01	±0.01	2.71	±0.04
	YR	1.01	±0.01	2.70	±0.04

Table S5. Comparison of regression techniques using MT data generation scheme with 5000 runs and linear measurement uncertainty following Eqs. (23) & (24). Unbiased (difference <5% to true value) slopes and intercepts are highlighted with grey background.

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 13 Slope=4, Intercept=0 γ_{Unc}=30% R^2 (POC,EC) = 0.76±0.01	OLS	3.48	±0.04	2.87	±0.18
	DR λ=1	4.53	±0.05	-2.94	±0.24
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	4.00	±0.05	-0.00	±0.22
	ODR	4.53	±0.05	-2.94	±0.24
	WODR	4.00	±0.03	-0.00	±0.09
	YR	4.00	±0.03	-0.00	±0.09
Case 14 Slope=4, Intercept=3 γ_{Unc}=30% R^2 (POC,EC) = 0.73±0.01	OLS	3.48	±0.04	5.87	±0.19
	DR λ=1	4.67	±0.05	-0.67	±0.26
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	3.98	±0.05	3.08	±0.23
	ODR	4.67	±0.05	-0.67	±0.26
	WODR	4.02	±0.03	2.68	±0.11
	YR	4.02	±0.03	2.68	±0.11

Table S6. Comparison of regression techniques using MT data generation scheme with 5000 runs. Linear measurement uncertainty follows Eqs. (23) & (24).

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 15 Slope=0.5 Intercept=0 γ_{Unc}=30% R^2 (POC,EC) = 0.54±0.01	OLS	0.40	±0.01	0.55	±0.03
	DR λ=1	0.45	±0.01	0.26	±0.03
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	0.50	±0.01	0.01	±0.03
	ODR	0.45	±0.01	0.26	±0.03
	WODR	0.52	±0.01	-0.23	±0.02
	YR	0.52	±0.01	-0.23	±0.02
Case 16 Slope=0.5 Intercept=3 γ_{Unc}=30% R^2 (POC,EC) = 0.40±0.01	OLS	0.40	±0.01	3.54	±0.04
	DR λ=1	0.50	±0.01	2.98	±0.04
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	0.50	±0.01	3.00	±0.04
	ODR	0.50	±0.01	2.98	±0.04
	WODR	0.52	±0.01	2.65	±0.04
	YR	0.52	±0.01	2.65	±0.04

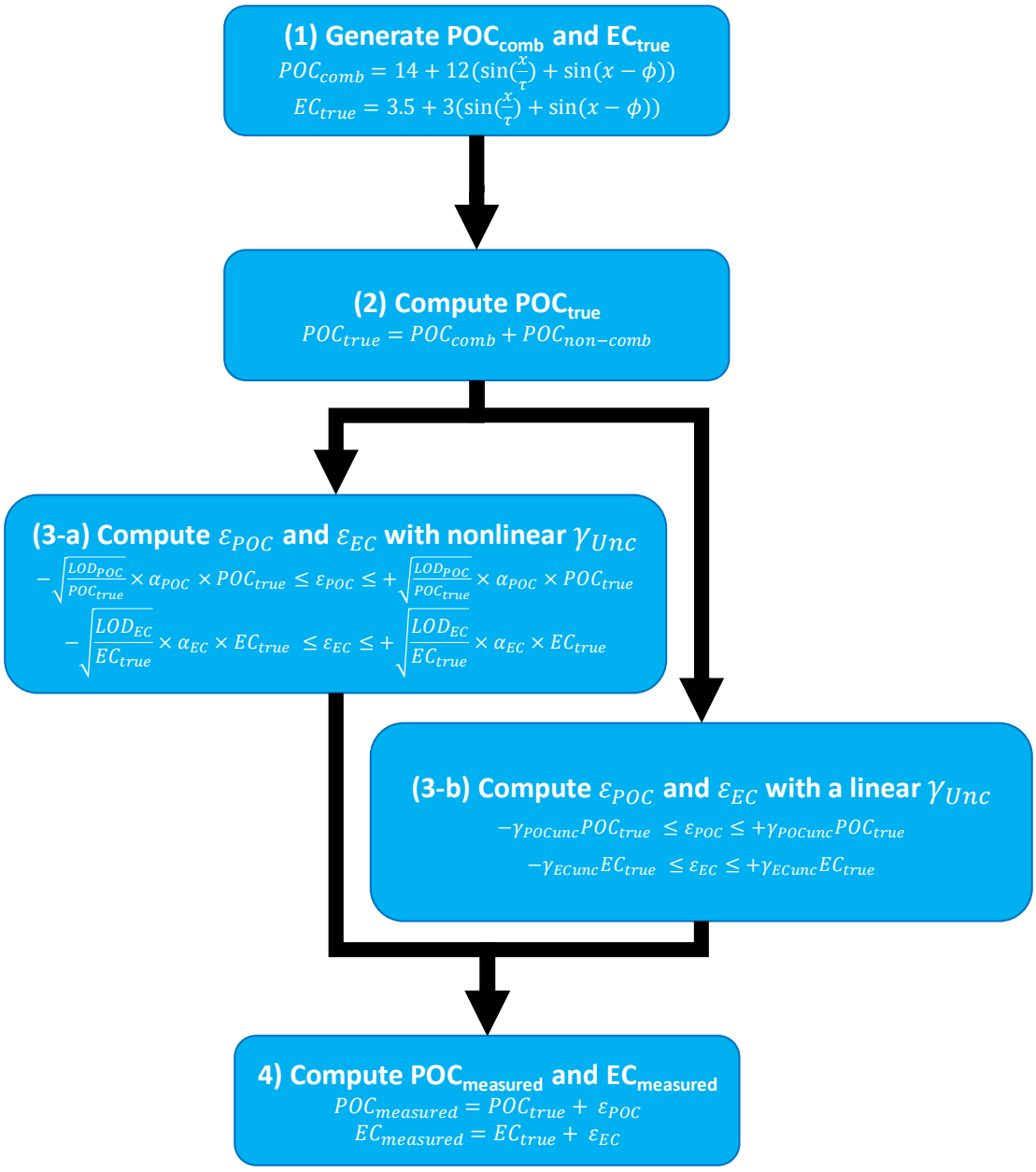
Table S7. Comparison of regression techniques using MT data generation scheme with 5000 runs. Linear measurement uncertainty follows Eqs. (23)&(24).

Case	Regression scenario	Slope mean	Slope SD	Intercept mean	Intercept SD
Case 17 Slope=1 Intercept=0 γ_{Unc}=30% R^2 (POC,EC) = 0.65±0.01	OLS	0.80	±0.01	1.07	±0.04
	DR λ=1	1.00	±0.01	0.00	±0.05
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	1.00	±0.01	0.00	±0.05
	ODR	1.00	±0.01	0.00	±0.05
	WODR	1.00	±0.01	0.00	±0.04
	YR	1.00	±0.01	0.00	±0.04
Case 18 Slope=1 Intercept=3 γ_{Unc}=30% R^2 (POC,EC) = 0.59±0.01	OLS	0.80	±0.01	4.07	±0.05
	DR λ=1	1.07	±0.01	2.62	±0.07
	DR $\lambda = \frac{\omega(X_i)}{\omega(Y_i)}$	1.00	±0.01	3.00	±0.06
	ODR	1.07	±0.01	2.62	±0.07
	WODR	1.02	±0.01	2.84	±0.05
	YR	1.02	±0.01	2.84	±0.05

127 Table S8. Abbreviations used for 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
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
γ_{RSD}	the ratio between the RSD values of $(OC/EC)_{pri}$ and EC

128



129

130 Figure S1. Flowchart of data generation steps using the sine functions of Chu (2005).

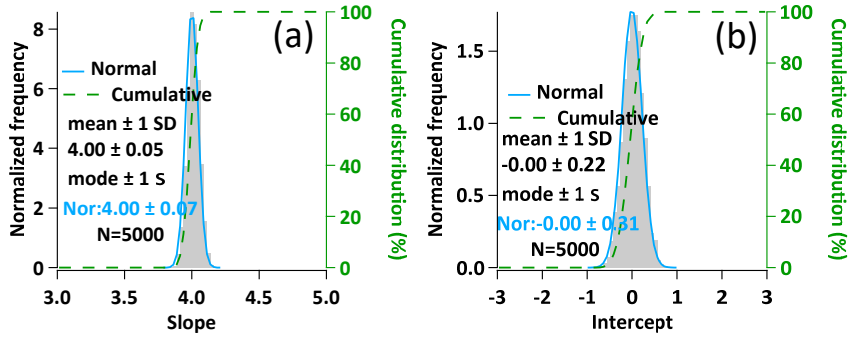
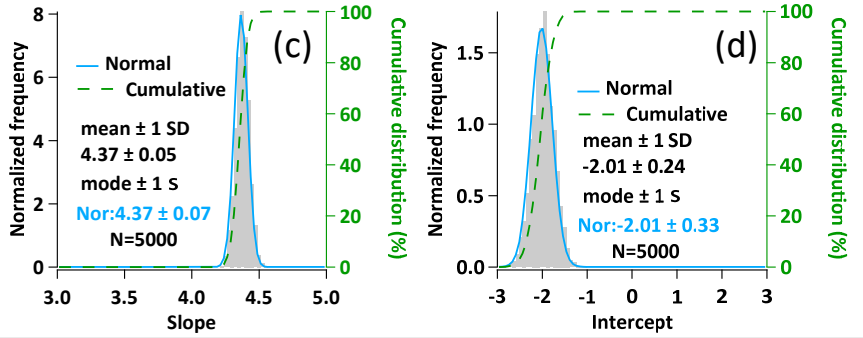
γ_{Unc}=30%	Slope=4	Intercept=0
Correct λ ($\lambda = \frac{POC^2}{EC^2}$)		
Incorrect λ ($\lambda = \frac{POC}{EC}$)		

Figure S2. 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} ($\lambda = \frac{POC}{EC}$).

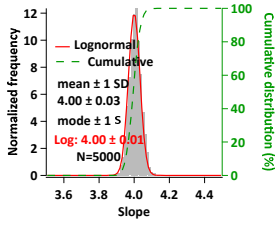
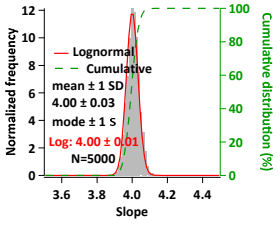
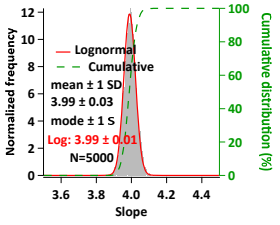
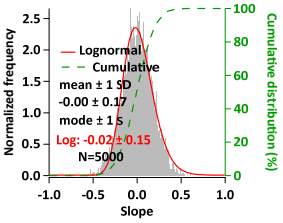
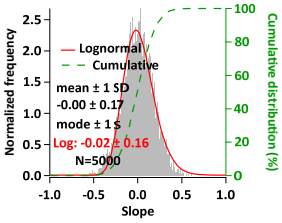
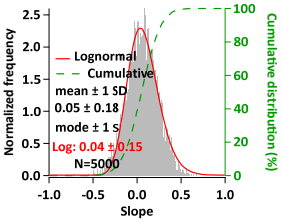
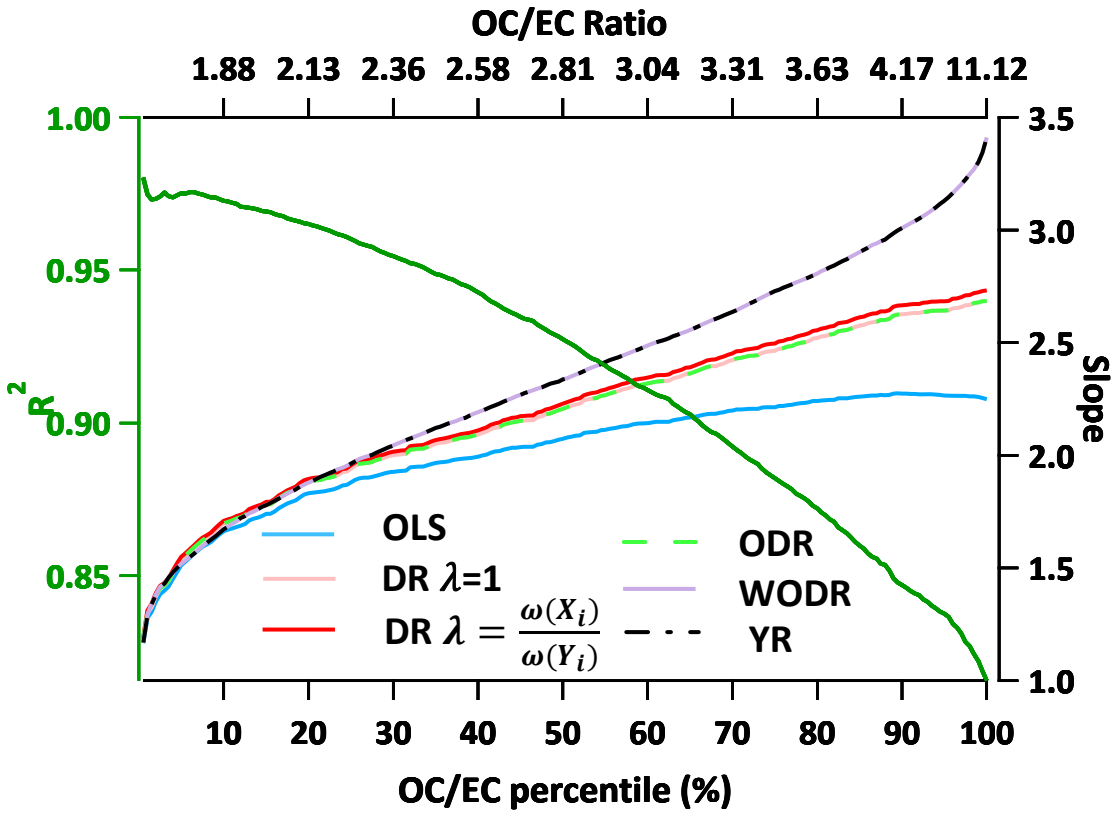
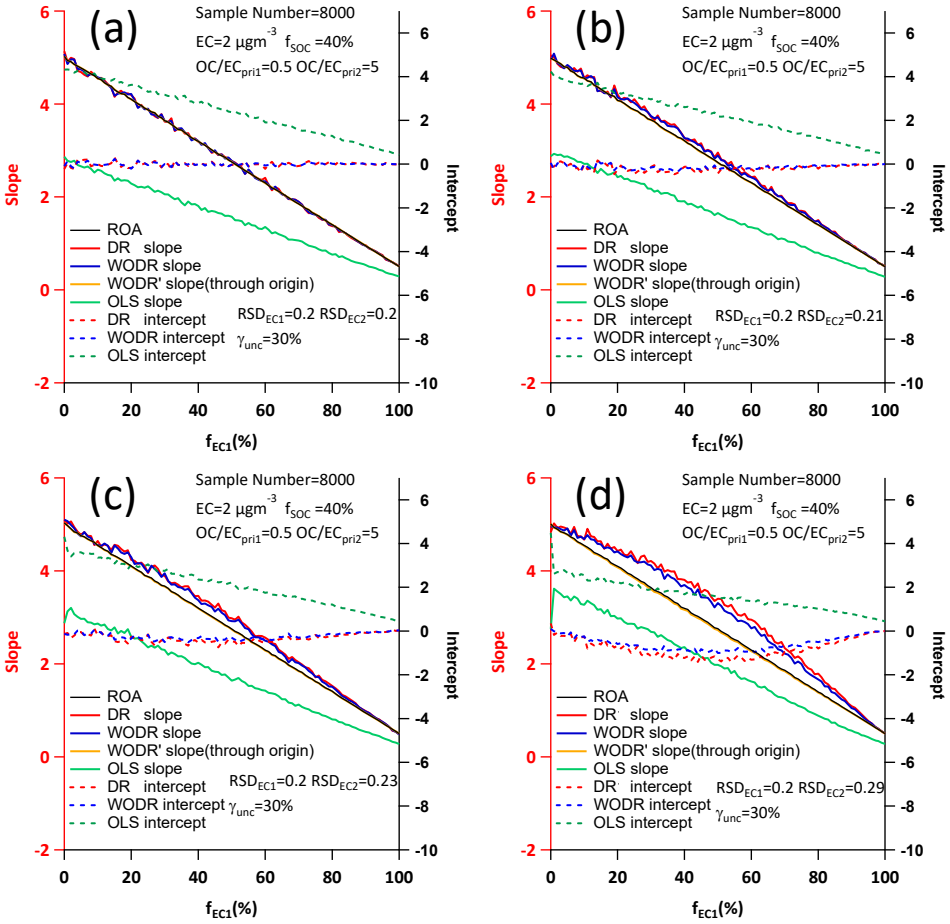
	λ by mean	λ by median	λ by mode
Slope=4			
Intercept=0			

Figure S3. Sensitivity tests of λ calculated by mean, median and mode.



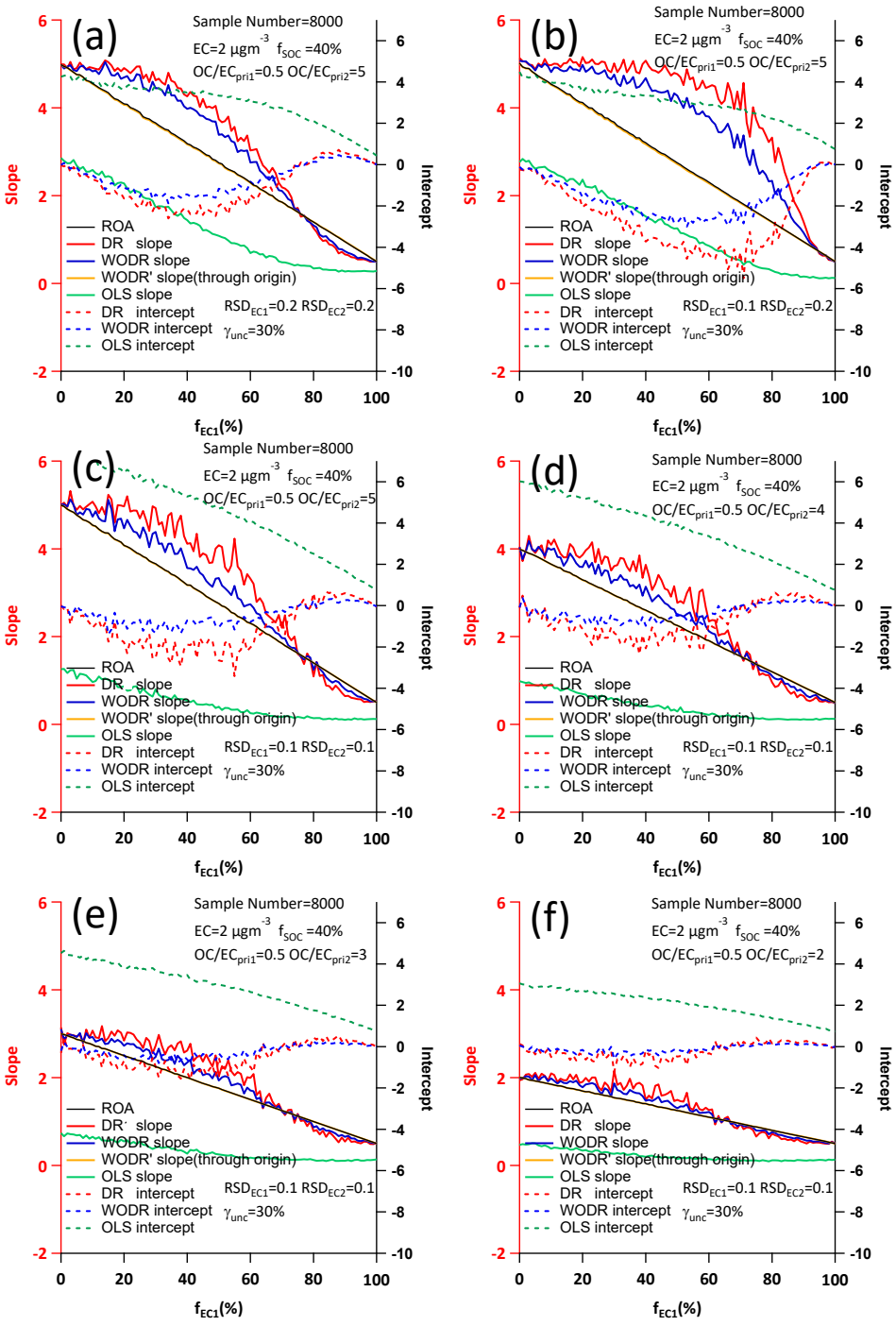
142

143 Figure S4. Regression slopes as a function of OC/EC percentile. OC/EC percentile range
144 from 0.5% to 100%, with an interval of 0.5%.



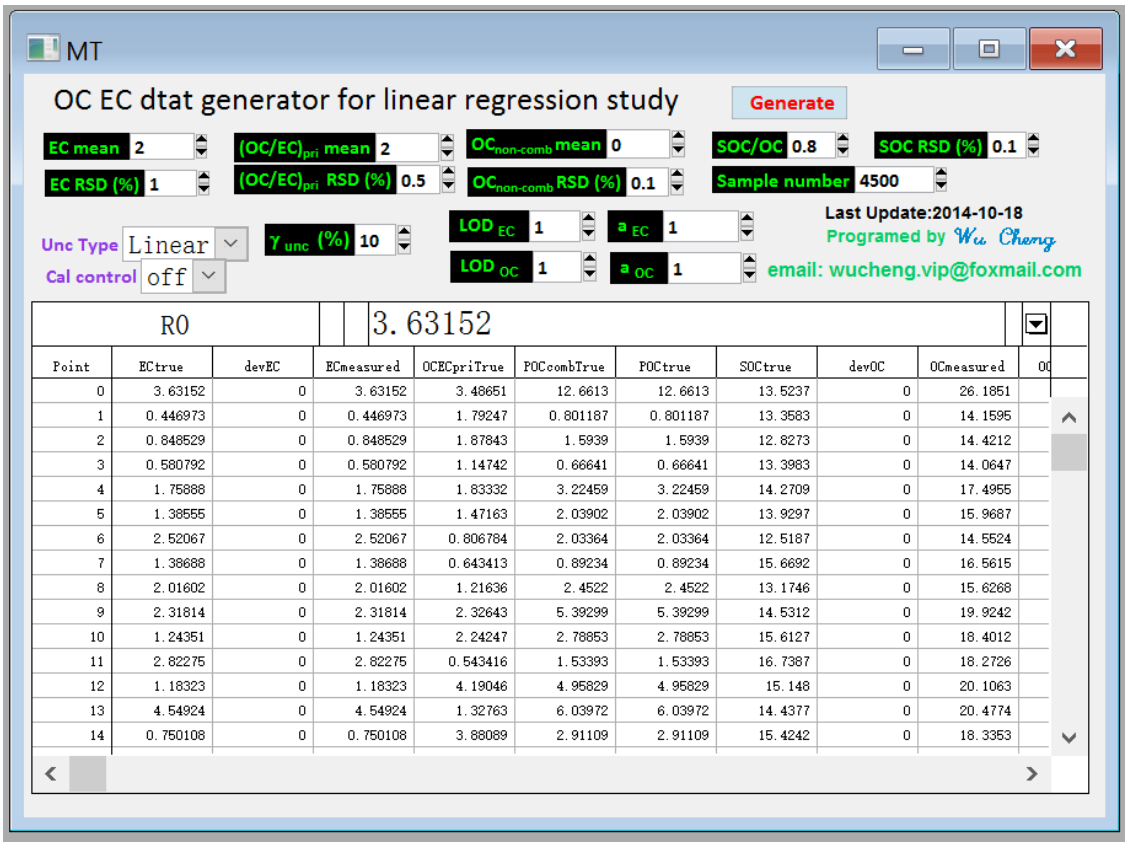
145

146 Figure S5. Study of two correlated sources secnario by different R^2 between the two
147 sources. (a) $R^2 = 1$ (b) $R^2 = 0.86$ (c) $R^2 = 0.75$ (d) $R^2 = 0.49$



148

149 Figure S6. Study of two independent sources secnario by different parameters. (a)γ_pri=10,
150 RSD_{EC1}=0.2,RSD_{EC2}=0.2 (b) γ_pri=10, RSD_{EC1}=0.1,RSD_{EC2}=0.2 (c) γ_pri=10,
151 RSD_{EC1}=0.1, RSD_{EC2}=0.1 (d) γ_pri=8, RSD_{EC1}=0.1, RSD_{EC2}=0.1(e) γ_pri=6, RSD_{EC1}=0.1,
152 RSD_{EC2}=0.1 (f) γ_pri=4, RSD_{EC1}=0.1, RSD_{EC2}=0.1



Point	ECtrue	devEC	ECmeasured	OCECpriTrue	POCcombTrue	POCtrue	SOCtrue	devOC	OCmeasured	OC
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 S7. 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.