Journal of Physics:
Conference Series

OPEN ACCESS

Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science

To cite this article: Ling Leng *et al* 2007 *J. Phys.: Conf. Ser.* **78** 012084

View the article online for updates and enhancements.

You may also like

- AOR: adaptive orthogonal regression for depression feature selection with incomplete EEG channels
Wenjia Dong, Zhijian Gong, Xueyuan Xu et al.
- Refereeing physics papers
C N Davies and J P Killingbeck
- Air ionizers case study
Cheng Qian



This content was downloaded from IP address 216.73.216.56 on 19/11/2026 at 08:52

Ordinary Least Square Regression, Orthogonal Regression, Geometric Mean Regression and their Applications in Aerosol Science

Ling Leng[1], Tianyi Zhang[1], Lawrence Kleinman [2], Wei Zhu[1]

Production Editor, *Journal of Physics: Conference Series*,
[1]Department of Applied Math and Statistics, Stony Brook, NY, 11790
[2]Department of Environmental Sciences, Brookhaven National Laboratory, NY, 11973

E-mail: weizhu@notes.cc.sunysb.edu

Abstract. Regression analysis, especially the ordinary least squares method which assumes that errors are confined to the dependent variable, has seen a fair share of its applications in aerosol science. The ordinary least squares approach, however, could be problematic due to the fact that atmospheric data often does not lend itself to calling one variable independent and the other dependent. Errors often exist for both measurements. In this work, we examine two regression approaches available to accommodate this situation. They are orthogonal regression and geometric mean regression. Comparisons are made theoretically as well as numerically through an aerosol study examining whether the ratio of organic aerosol to CO would change with age.

1. Introduction and a General Structural Model

The classical ordinary least squares regression theory relies on the assumption that the explanatory variables are measured without error. In aerosol science as well as in many other scientific disciplines, this assumption is often found untrue when randomness exists in the regressors due to measurement error or other underlying volatility. Two popular approaches, the orthogonal regression and the geometric mean regression, have been proposed for the analysis when both the dependent and the independent variables are random. The immediate question confronting the scientist is which approach to adopt for his or her data, and what to do if neither approach is suitable. In this paper, we address this question in the context of simple linear regression through the analysis of a general structural model suitable when both variables are random.

Suppose both X and Y contain some random errors, δ and ε, which may come from measurement or other resources. A suitable model is as follows.

$$X = \xi + \delta \qquad \delta \sim N\left(0, \sigma_\delta^2\right)$$
$$Y = \eta + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$$
$$\eta = \beta_0 + \beta_1 \xi,$$

© 2007 IOP Publishing Ltd

where δ and ε are independent random errors. There are two analysis approaches concerning this model: the functional and the structural. The basic difference between the two approaches is whether to consider ξ as a non-random variable or a random variable following normal distribution with mean μ and variance τ^2, and independent to both random errors. Since the latter approach is more general, in the discussion below, we will follow the structural model where X and Y follow a bivariate normal distribution with mean and covariance structure as follows:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \beta_0 + \beta_1 \mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma_\delta^2 & \beta_1 \tau^2 \\ \beta_1 \tau^2 & \beta_1^2 \tau^2 + \sigma_\varepsilon^2 \end{pmatrix} \right)$$

2. Comparing Different Regression Approaches

2.1. Estimation Based on the General Structural Model

For the general structural model above, its mean vector and covariance matrix can be easily derived as follows:

$$E(X) = E(\xi) + E(\delta) = \mu$$

$$E(Y) = E(\eta) + E(\varepsilon) = \beta_0 + \beta_1 E(\xi) = \beta_0 + \beta_1 \mu$$

$$\mathrm{var}(X) = \mathrm{var}(\xi) + \mathrm{var}(\delta) = \tau^2 + \sigma_\delta^2$$

$$\mathrm{var}(Y) = \mathrm{var}(\eta) + \mathrm{var}(\varepsilon) = \mathrm{var}(\beta_1 \xi) + \mathrm{var}(\varepsilon) = \beta_1^2 \tau^2 + \sigma_\varepsilon^2$$

$$\mathrm{cov}(X, Y) = \mathrm{cov}(\xi + \delta, \beta_0 + \beta_1 \xi + \varepsilon) = \beta_1 \tau^2$$

Given a random sample of observed X's and Y's, we can obtain the MLE of the slope of the regression. Its value, however, depends on the ratio of the two error variances $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2$ [1]

$$\hat{\beta}_1 = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2}}{2S_{XY}} \quad [2]$$

Inference (hypothesis test, confidence interval) on the slope parameter can be carried out similarly using the maximum likelihood approach. We consider this the general and correct approach when both variables are random. Since it is a parametric model, the readers are reminded that normality transformation should be performed prior to the regression analysis if a variable is found not normal. In the following, we will compare two commonly used regression methods when both X and Y are random, the orthogonal regression and the geometric mean regression, to this general approach. Guideline will be provided on whether and when each approach is considered suitable.

2.2. Ordinary Least Square Regression (OLS)

As illustrated in Figures 1a and 1b, the ordinary least square (OLS) estimate of Y on X will minimize the squared vertical distance $\sum (y_i - \beta_0 - \beta_1 x_i)^2$ from the points to the regression line. The OLS estimate of the slope is $\hat{\beta}_1 = S_{XY} / S_{XX}$. This is the case when $\lambda = \infty$ in the general structural modelling approach. Similarly, the OLS estimate of X on Y would minimize the horizontal distance to the regression line. The latter is also called the reverse regression. The OLS is suitable when only one of the two variables is random.

2.3. Orthogonal Regression (OR)

Instead of minimizing the vertical (or horizontal) distance as in the OLS, the orthogonal regression takes the middle ground by minimizing the orthogonal distance from the observed data points to the regression line as illustrated in Figure 1c.

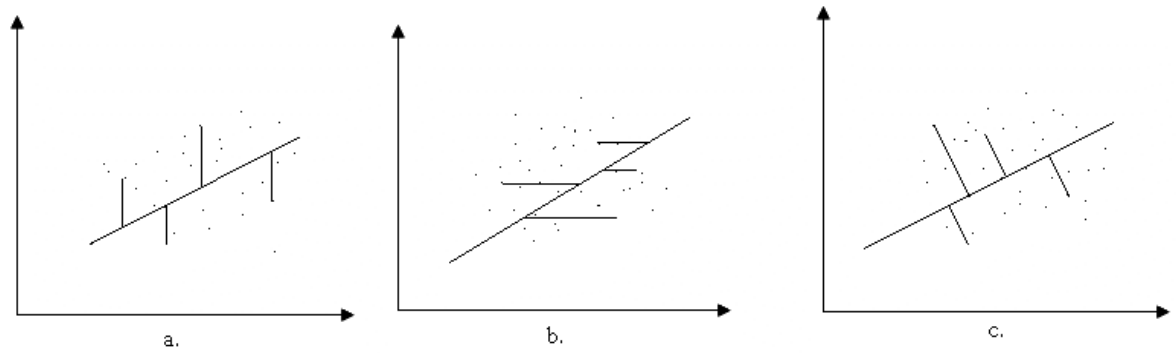


Figure 1. The ordinary least squares regression (a, b) and the orthogonal regression (c).

2.3.3. *Point Estimate*

The resulting OR estimate of β_1 is:

$$\hat{\beta}_1 = \frac{\frac{S_{YY} - S_{XX}}{S_{XY}} + \sqrt{\Delta}}{2} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}{}^2}}{2S_{XY}}$$

This is the same as the MLE in the general structural modelling approach when $\lambda = 1$.It means that the orthogonal regression is suitable when the error variances are equal.

2.3.2 *The Connection Between Orthogonal Regression and PCA*

There is a close relationship between the Principle Component Analysis (PCA) and the Orthogonal Regression [3].

For the sample covariance matrix of the random variables (X, Y), $\begin{bmatrix} s_{XX} & s_{XY} \\ s_{XY} & s_{YY} \end{bmatrix}$, its highest eigenvalue is $\lambda = \frac{s_{XX} + s_{YY} + \sqrt{(s_{XX} + s_{YY})^2 - 4(s_{XX}s_{YY} - s_{XY}^2)}}{2}$. And the eigenvector (first principal component) corresponding to this eigenvalue is $\left(s_{XY}, \frac{s_{YY} - s_{XX} + \sqrt{(s_{XX} + s_{YY})^2 - 4(s_{XX}s_{YY} - s_{XY}^2)}}{2} \right)$.Therefore, the slope of the first principal component is $\frac{s_{YY} - s_{XX} + \sqrt{(s_{XX} + s_{YY})^2 - 4(s_{XX}s_{YY} - s_{XY}^2)}}{2s_{XY}}$, which is the same as the slope estimator from the orthogonal regression.

Intuitively, the first principal component is the line passing through the greatest dimension of the concentration ellipse, which coincides with the orthogonal regression line. Therefore, existing statistical inference techniques for the PCA can be applied directly to the inference of the slope parameter, β_1, from the OR approach as shown in the following.

2.3.3. *Inference for Orthogonal Regression*

From its equivalence with PCA, we can obtain the confidence interval and conduct hypothesis for the orthogonal regression slope as follows.

Let l_1 and l_2 be the eigenvalues of the sample covariance matrix ($l_1 > l_2$), $\hat{\theta} = \tan^{-1}(\hat{\beta}_1)$ and $\Phi_L = \sin^{-1}\left[\chi^2_\alpha(1)/\left\{(n-1)\left[\frac{l_1}{l_2}+\frac{l_2}{l_1}-2\right]\right\}\right]^{1/2}$, the 100(1-α)% large sample CI for the slope is:

$$\tan\left(\hat{\theta}-\Phi_L\right) \le \beta_1 \le \tan\left(\hat{\theta}+\Phi_L\right).$$

Similarly, we obtain the following hypothesis test for the slope: $H_0: \beta_1 = \beta_{10}$ *v.s.* $H_a: \beta_1 \neq \beta_{10}$

if $[(n-2)r^2/(1-r^2)] \le F_a(1,n-2)$, *accept* $H_0: \beta_1 = \beta_{10}$;

if $[(n-2)r^2/(1-r^2)] > F_a(1,n-2)$, *reject* $H_0: \beta_1 = \beta_{10}$

where $r^2 = (l_1-l_2)^2/(l_1+l_2)^2$

2.4. The Geometric Mean Regression (GMR)

Besides the orthogonal regression, another intuitive approach of taking the middle ground when both X and Y are random is to simply take the geometric mean of the slope of y on x regression line, and the reciprocal of the slope of x on y regression line. This approach is called the "geometric mean regression" (GMR). By definition, the estimated slope via the GMR approach is

$$\hat{\beta}_1 = sign(S_{XY})\sqrt{\hat{\beta}_{OLS,\ \text{Y on X}} * (\hat{\beta}_{OLS,\ \text{X on Y}})^{-1}}$$

Simplification of the above formula yields: $\hat{\beta}_1 = sign(S_{XY})\sqrt{\frac{S_{YY}}{S_{XX}}}$

Comparing to the MLE in the bivariate normal structural modelling approach, we notice that the GMR estimator is equal to the MLE if and only if $\lambda = S_{YY}/S_{XX}$ [4].This means that the GMR approach is suitable when the randomness from X and Y are from the random errors only. That is, when we take the functional analysis approach by assuming that ξ is not random.

3. Data Analysis

We measured organic aerosol and CO concentration at 10 different ages (time since the CO was emitted) to examine if the ratio of organic aerosol to CO would change with age. On the time scale of interest CO is inert so this change would reflect the atmospheric reactions forming organic aerosols [8].

We found it plausible to assume that the log transformation of CO follows normal distribution [9] [10], and the error variances for both measurements are equal [8]. That is $\lambda = 1$. This means OR, coinciding with the general structural modeling approach, is presumably the most suitable model to use in our case.

For comparison purposes, we also examined the OLS and the GMR models at each age point. It can be shown theoretically that the MLE of slope will decrease while λ increases:

$$\frac{\partial\hat{\beta}_1}{\partial\lambda} = \frac{1}{2S_{XY}}\{-S_{XY} + \frac{1}{2}[(S_{YY}-\lambda S_{XX})^2 + 4\lambda S_{XY}^2]^{-\frac{1}{2}} * [4S_{XY}^2 - 2(S_{YY}-\lambda S_{XX})S_{XX}]\}$$

$$\le \frac{1}{2S_{XY}}\{-S_{XY} + \frac{1}{2}[(S_{YY}-\lambda S_{XX})^2 + 4\lambda S_{XX}S_{YY}]^{-\frac{1}{2}} * [4S_{XX}S_{YY} - 2(S_{YY}-\lambda S_{XX})S_{XX}]\}$$

$$(\text{using } S_{XY}^2 \le S_{XX}S_{YY}) \le \frac{1}{2S_{XY}}[-S_{XY} + \frac{1}{2}(S_{YY}+\lambda S_{XX})^{-1} * (2S_{XX}S_{YY} + 2\lambda S_{XX}^2)] \le 0$$

This is reflected from the analysis of our particular data set. An obvious descending trend is shown from OR $(\lambda = 1)$ to GMR (λ ranges from 30 to 140) to OLS $(\lambda = \infty)$ at each time point (Figure 2).

In summary, we found that the estimated regression lines using the GMR and OLS approaches both fell out of the 95% confidence interval (CI) of orthogonal regression estimate although the geometric mean regression provided closer estimates to the OR than the OLS. This illustrates that different regression approaches can yield drastically different results for a given data set.

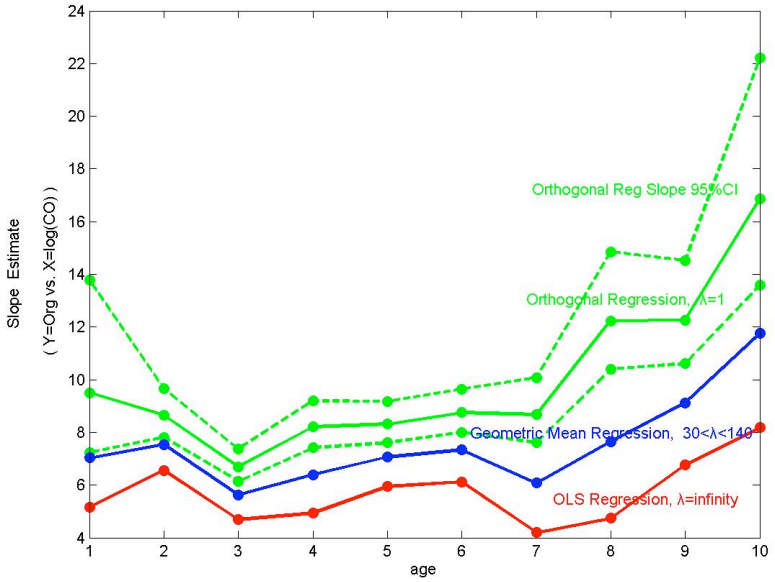


Figure 2. Analysis of data from an aerosol study

References

[1] P. Sprent. Models in Regression and Related Topics. *Methuen's Statistical Monographs*. Matheun & Co Ltd, London, 1969.

[2] M.Y. Wong. Likelihood estimation of a simple linear regression model when both variables have error. *Biometrica,* 76(1): 141-148,1989.

[3] J. D. Jackson and J. A. Dunlevy. Orthogonal Least Squares and the Interchangeability of Alternative Proxy Variables in the Social Sciences. *The Statistician*, 37(1): 7-14, 1988.

[4] P. Sprent and G. R. Dolby. Query: the geometric mean functional relationship. *Biometrics*, 36(3): 547-550, 1980.

[5] P. Jolicouer. Linear regressions in Fishery research: some comments. *J. Fish. Res. Board Can.*, 32(8): 1491-1494, 1975.

[6] M. A. Creasy. Con_dence limits for the gradient in the linear functional relationship. *J. Roy. Statist. Soc. Ser. B.*, 18:65-69, 1956.

[7] F. Barker, Y. C. Soh, and R. J. Evans. Properties of the geometric mean functional relationship. *Biometrics*, 44(1): 279-281, 1988.

[8] L. Kleinman, et al. Time Evolution of Aerosol Composition over the Mexico City Plateau (manuscript in preperation for Atmospheric Chemistry and Physics Discussions), 2007.

[9] R. McGraw and R. Zhang. Multivariate analysis of homogeneous nucleation rate measurements:I. Nucleation in the p-toluic acid/sulfuric acid/water system. (manuscript), 2007.