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]

[1]Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794

[2] Department of Environmental Sciences, Brookhaven National Laboratory, NY 11973

Introduction

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 CO to organic aerosol would change with age.

1. The Bivariate Normal Structural Model

- **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.**

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

where ξ, δ and ε are independent.

- **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. Since the latter is more general, in the discussion below, we will focus on the structural model.**
- **Thus, we assume that X and Y follow a bivariate normal distribution:**

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

2. Maximum Likelihood Estimator (MLE)

- **The mean vector and covariance matrix of X and Y can be easily derived as follows:**

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

- **Subsequently, we can obtain the maximum likelihood estimator 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].**
- **When $\lambda = 1$, the MLE of β_1 is:**

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

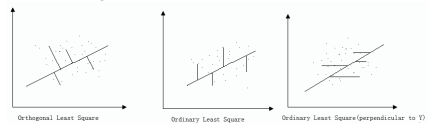
- **When $\lambda = \lambda_0$, the MLE of β_1 is:**

$$\hat{\beta}_1 = \frac{S_{YY} - \lambda_0 S_{XX} + \sqrt{(S_{YY} - \lambda_0 S_{XX})^2 + 4\lambda_0 S_{XY}^2}}{2S_{XY}} \quad [2]$$

- **Inference (hypothesis test, confidence interval) on the slope parameter can be carried out similarly using the maximum likelihood approach. We consider this the correct approach. 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.**

3. Orthogonal Regression (OR)

- **As illustrated in the following plots, The ordinary least square (OLS) estimate (of Y on x) will minimize the squared vertical distance from the points to the regression line. And the estimate of the slope is: $\hat{\beta}_1 = S_{XY}/S_{XX}$ (this is the case when $\lambda = \infty$ in the general modeling approach). Alternatively, The OLS estimate (of X on Y) would minimize the horizontal distance to the regression line. The orthogonal regression takes the middle ground by minimizing the orthogonal distance to the regression line.**



- **The resulting 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 bivariate normal structural modeling approach when $\lambda = 1$.**
- **<u>This means that the orthogonal regression is suitable when the error variances are equal.</u>**

4. Connection between OR & PCA

- **There is a close relationship between the Principle Component Analysis (PCA) and the orthogonal regression (OR).**
- **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:**

$$\lambda = \frac{S_{XX} + S_{YY} + \sqrt{(S_{XX} + S_{YY})^2 - 4(S_{XX}S_{YY} - S_{XY}^2)}}{2}$$

The eigenvector corresponding to this eigenvalue is:

$$\left(S_{XX}, \frac{S_{YY} - S_{XX} + \sqrt{(S_{XX} - S_{YY})^2 + 4S_{XY}^2}}{2} \right)$$

Therefore, the slope of the 1st PC is: $\frac{S_{YY} - S_{XX} + \sqrt{(S_{XX} - S_{YY})^2 + 4S_{XY}^2}}{2S_{XY}}$ 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 principal component analysis can be applied directly to inference of the slope parameter, β_1, from the orthogonal regression approach as shown in the following slide.**

5. Inference on OR

- **Confidence Interval for β_1 [3]: Let l_1 and l_2 be the eigenvalues of the sample covariance matrix ($l_1 > l_2$),**

$$\hat{\theta} = \tan^{-1}(\hat{\beta}_1), \ \ and \ \ \Phi_L = \sin^{-1}\left[\chi_\alpha^2(1) / \left\{ (n-1)\left[\frac{l_1}{l_2} + \frac{l_2}{l_1} - 2 \right] \right\} \right]^{1/2},$$

the 100(1- α)% large sample CI for the slope is

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

- **Hypothesis Test of** $H_0 : \beta_1 = \beta_{1_0}$, $H_a : \beta_1 \ne \beta_{1_0}$

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

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

where $r^2 = \frac{(l_1 - l_2)^2}{(l_1 + l_2)^2}$

6. Geometric Mean Regression

- **Another intuitive approach of taking the middle ground when both X and Y are considered 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.**

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

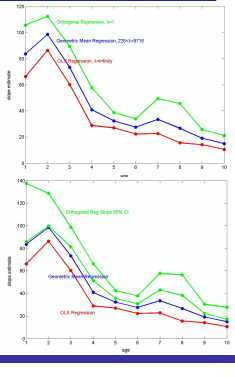
- **Simplification of the above formula yields:**

$$\hat{\beta}_1 = sign(S_{XY}) \sqrt{\frac{S_{YY}}{S_{XX}}}$$

- **Comparing to the MLE, we notice that the geometric mean estimator is equal to the MLE if and only if $\lambda = S_{YY}/S_{XX}$ [4].**
- **<u>This means that the geometric mean approach is suitable when the randomness from X and Y are from the random errors only.</u> That is, when we take the functional analysis approach by assuming that ξ is not random.**
- **It yields a biased estimate unless : $\lambda = \beta^2 = \sigma_\varepsilon^2 / \sigma_\delta^2$ [5]**
- **Barker et al (1988) [6] presented a "least triangles approach". They defined a new loss function to obtain the regression estimate. This regression line would minimize the sum of the areas of the right-angled triangles formed from the data point and the regression line.**

7. Data Analysis

- **Our data are organic aerosol and CO concentration measured at 10 different ages (time since the CO was emitted). We are examining if the ratio of CO to organic aerosol changes with age. On the time scale of interest CO is inert so this change would be due to the atmospheric reactions forming organic aerosols. [7]**
- **We found it reasonable to assume equal error variances for both measurements [7]. That is λ =1. This means the orthogonal regression (OR), coinciding with the bivariate normal structural model approach in our case, is the most suitable model to use.**
- **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. This is reflected from the analysis of this particular data set. An obvious descending tendency is shown from OR to GMR to OLS.**
- **The estimated regression lines using the GMR and OLS approaches fell out of the 95 CI of orthogonal regression estimate. The geometric mean regression provided closer estimates to the OR than the OLS.**



8. Conclusion

- **In this work we presented a general bivariate normal structural model and inference framework suitable when both X and Y are random in a regression scenario. It is shown that the MLE of the slope parameter depends on λ, the ratio of the error variances of Y and X.**
- **It is shown that ordinary least squares (OLS), orthogonal regression (OR), geometric mean regression (GMR) can all be considered as special cases of the above general model with λ = ∞, 1, and S_{YY}/S_{XX} respectively.**
- **Subsequently, we conclude that the (1) the orthogonal regression is suitable when the error variances are equal, and (2) the geometric mean approach is suitable when the randomness from X and Y are from the random errors only.**
- **We found that the orthogonal regression provides the right model for the estimation of CO to organic aerosol concentration ratio and its temporal trend. Comparison to the results from the orthogonal regression model, we found that both the GMR and the OLS yielded severely downward biased estimates with the OLS being more biased than the GMR.**

9. Reference

- **[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] F. Barker, Y. C. Soh, and R. J. Evans. Properties of the geometric mean functional relationship. *Biometrics*, 44(1): 279-281, 1988.**
- **[7] L. Kleinman, et al. Time Evolution of Aerosol Composition over the Mexico City Plateau (manuscript in preperation for Atmospheric Chemistry and Physics Discussions), 2007.**

BROOKHAVEN NATIONAL LABORATORY

Supported by

The U. S. Department of Energy

STONY BROOK STATE UNIVERSITY OF NEW YORK