LEAST SQUARES REGRESSION WITH UNCERTAINTIES IN BOTH VARIABLES: A REALITY IN ANALYTICAL CHEMISTRY

Elcio Cruz de Oliveira [1], *Paula Fernandes de Aguiar* [2]

[1] PETROBRAS TRANSPORTE S.A., Rio de Janeiro, Brazil, elciooliveira@petrobras.com.br
[2] Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, paulafda@iq.ufrj.br

Abstract: The least squares regression of data with error in *x* and *y* should not be implemented by ordinary least squares (OLS). In this work, it is discussed orthogonal distance regression (ODR) as an alternative approach in order to take into account the uncertainty in *x* variable. The first example, comparison of two methods, shows that ODR technique leads to a difference conclusion than t-test. The second case study, an analytical curve, shows that when the variance of the replicates in a single *x* value is tiny, when compared with the variance from *x* variable, there is no significant difference between ODR and OLS coefficients because the uncertainty is negligible in *x*-axis.

Key words: orthogonal distance regression, least squares regression, uncertainty in both variables, equivalence of the methods.

1. INTRODUCTION

Classical univariate regression is the most used regression method in Analytical Chemistry. It is generally implemented by ordinary least squares (OLS) fitting, using n points (x_i, y_i) to a response function, usually linear [1] and handling homoscedastic data. In this way, it is estimated the amount of the unknown (x_0) from one or more measurements of its response (y_0). The algorithms for carrying out such analytical curve have been well established in the literature. When the data are heteroscedastic, the Analytical Chemistry uses weighted linear regression.

But a problem remains in the analytical community: the uncertainty in *x*-axis. Classic linear regression, available in commercial softwares, assumes that *x* variable errors are negligible, that is, error-free [1-2].

As analytical methods usually have to be applicable over a wide range of concentrations, a new method is often compared with a standard method by analysis of samples in which the analyte concentration may vary over several powers of ten. In this case, it is inappropriate to use the paired *t*-test since its validity rests on the assumption that any errors, either random or systematic, are independent of concentration [3]. Over wide ranges of concentration this assumption may no longer be true. A second problem appears when, certified reference materials are not available, and that usually has negligible uncertainty, in order to carry out an analytical curve. So, beyond the uncertainty derived from the signal, it must be considered the uncertainty from *x*-axis. In these cases, OLS should not be used, so the literature suggests carrying out orthogonal distance regression (ODR). The aim of this work is to suggest how to handle these cases when uncertainties in both variables are considered.

2. METHODOLOGY

2.1. General

Generally, it is assumed that only the response variables, y, is subject to error and that the predictor variable, x, is known with negligible error. However, there are situations for which the assumption that x is error free is not justified. In these situations, it is required regression methods that take the error in both variables into account. They are called errors-in-variables regression methods.

If η_i represents the true value of y_i and ξ_i the true value of x_i, with ε_i and δ_i the experimental errors, respectively, then: $y_i = b_0 + b_1 x_i + (\varepsilon_i - b_1 \delta_1)$, where the last term represents the experimental errors. The fitted line is then the one for which the least sum of squared, $d_i s$, is obtained and the method has been called orthogonal distance regression (ODR). This is equivalent to finding the first principal component of a data set consisting of 2 variables and n samples [4].

2.2. ODR statistics

The expression of the function of the likelihood method, when it is considered *n* pairs of values (x_i, y_i) and a multidimensional model suitable to describe experimental data fluctuations, is the multivariate normal [5]:

$$L(\alpha, \beta, \mu_{x_i} / x_i, y_i) = \prod_{i=1}^{n} \left\{ \frac{1}{\left(2\pi\sigma_{y_i}^2\right)^{1/2}} \times \frac{1}{\left(2\pi\sigma_{x_i}^2\right)^{1/2}} \right\} \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[\frac{(x_i - \mu_{x_i})^2}{\sigma_{x_i}^2} + \frac{(y_i - \alpha - \beta\mu_{x_i})^2}{\sigma_{y_i}^2} \right] \right\} \quad (1)$$

Both variables are affected by random measurement errors: x variable, $\sigma^2_{x_i} = \sigma^2_{\varepsilon_i}$ and y variable, $\sigma^2_{y_i} = \sigma^2_{\delta_i}$.

If it is considered that both variances of the variables are constant, $\sigma^2_x = \theta$ and its known rate is λ. This ratio can be defined as:

$$\lambda = \frac{\sigma^2_y}{\sigma^2_x} = \frac{\sigma^2_\delta}{\sigma^2_\varepsilon} = \frac{\lambda\theta}{\theta} \tag{2}$$

Applying (2) in (1):

$$L\left(\alpha, \beta, \mu_{x_i} / x_i, y_i\right) = \frac{1}{(2\pi\theta)^n \lambda^{n/2}} \exp\left\{-\frac{1}{2\lambda\theta}\sum_{i=1}^{n}\left[\lambda\left(x_i - \mu_{x_i}\right)^2 + \left(y_i - \alpha - \beta\mu_{x_i}\right)^2\right]\right\} \tag{3}$$

And its logarithm is given by:

$$l\left(\alpha, \beta, \mu_{x_i} / x_i, y_i\right) = -n\log(2\pi\theta) - \frac{n}{2}\log(\lambda) - \frac{1}{2\lambda\theta}\left[\sum_{i=1}^{n}\lambda\left(x_i - \mu_{x_i}\right)^2 + \sum_{i=1}^{n}\left(y_i - \alpha - \beta\mu_{x_i}\right)^2\right] \tag{4}$$

Maximizing (4), the log likelihood function, in relation to the disturbing parameters [6], $\hat{\mu}_{x_i}$:

$$\hat{\mu}_{x_i} = \frac{\lambda x_i + \beta(y_i - \alpha)}{\lambda + \beta^2} \tag{5}$$

Substituting equation (5) in equation (4), it is obtained the profiled log likelihood function that is function only of α, β and θ.

Deriving this new equation in relation of these three parameters and equalizing the derivatives to zero – the approach of Deming [7] estimates b_1, equation (6):

$$b_1 = \frac{s^2_y - \left(\frac{\sigma^2_\delta}{\sigma^2_\varepsilon}\right)\times s^2_x}{2\,\mathrm{cov}(y,x)} + \sqrt{\left(\frac{s^2_y - \left(\frac{\sigma^2_\delta}{\sigma^2_\varepsilon}\right)\times s^2_x}{2\,\mathrm{cov}(y,x)}\right)^2 + \left(\frac{\sigma^2_\delta}{\sigma^2_\varepsilon}\right)} \tag{6}$$

With s^2_y and s^2_x the variance of y variable and the x variable respectively;

$\mathrm{cov}(y,x) = \left(\sum (y_i - \bar{y})(x_i - \bar{x})\right) / (n-1)$ the covariance of y and x.

Since both variables are affected by random measurement errors and the simplest case is when $\sigma^2_\varepsilon = \sigma^2_\delta$, an unbiased estimation of the regression coefficients can be obtained by minimizing $\sum d_i^2$, i.e. the sum of the squares of the perpendicular distances from the data points to the regression line, where d_i is determined perpendicular to the estimated line [4].

The expression for b_1 and b_0 are:

$$b_1 = \frac{s^2_y - s^2_x + \sqrt{\left(s^2_x - s^2_y\right)^2 + 4(\mathrm{cov}(y,x))^2}}{2\,\mathrm{cov}(y,x)} \tag{7}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{8}$$

2.3. Confidence intervals

To test for bias, i.e. equivalence of the methods compared, the 95%-confidence intervals of the parameters from the linear equations $y = b_0 + b_1 x$ obtained after the orthogonal regression were used to test whether the optimal parameters of $b_0 = 0$ and $b_1 = 1$ are included in the spanned confidence intervals (CI) [8]:

$$CI(b_0) = b_0 \pm t_{P,f} \times s_{b_0} \text{ and } CI(b_1) = b_1 \pm t_{P,f} \times s_{b_1} \tag{9}$$

t = Student t-factor with: p = 95%; $f = n - 2$; s_{b_0} and s_{b_1} = standard deviation of the parameters b_0 and b_1.

The ideal values of a $b_0 = 0$ and $b_1 = 1$ imply no bias between the compared methods, i.e. equivalence in the calibration results. A fail of the test for the axis intercept b_0 imply a systematic bias, e.g. caused by a wrong blank correction of one method. If the test fails for the slope b_1, this implies a proportional bias. Combinations of the two errors can also appear.

2.4. OLS versus ODR

Mandel [9] considers an approximate relationship between the ordinary least squares slope, $b_1(OLS)$, and the orthogonal distance regression slope, $b_1(ODR)$ in (10):

$$b_1(ODR) = \frac{b_1(OLS)}{\left(1 - \frac{s^2_{ex}}{s^2_x}\right)} \tag{10}$$

Where s^2_{ex} is the variance of a single x value (involves replicate observations of the same x) and s^2_x is the variance of the x variable.

Table 1 shows the relation between s^2_{ex} and the ratio $b_1(ODR) / b_1(OLS)$, when a perfect system is considered, that is s^2_x is constant and equal to 1.

Table 1. Relation between s^2_{ex} / s^2_x and $b_1(ODR) / b_1(OLS)$

s^2_{ex} / s^2_x	$b_1(ODR) / b_1(OLS)$
0.00	1.00
0.01	1.01
0.10	1.11
0.20	1.25
0.25	1.33

0.33	1.50
0.50	2.00
0.70	3.33
0.90	10.0

Figure 1 shows that s_{ex}^2 / s_x^2 and $b_1(ODR) / b_1(OLS)$ have a behavior that is close to the linearity, when the variance of a single x value is lower than the variance of the x variable that is from 0.0 to 0.2.

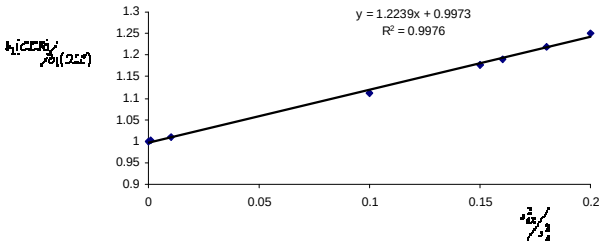

Fig. 1. Linear relation between s_{ex}^2 / s_x^2 and $b_1(ODR) / b_1(OLS)$

When the s_{ex}^2 / s_x^2 increases up to 0.5, the best regression seems to be quadratic, Figure 2.

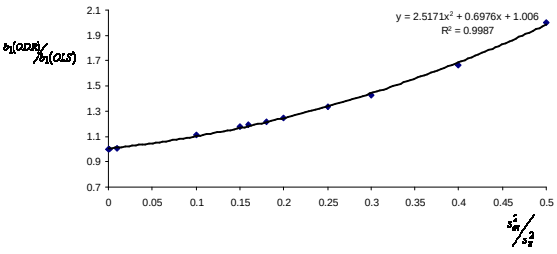

Fig. 2. Quadratic regression between s_{ex}^2 / s_x^2 and $b_1(ODR) / b_1(OLS)$

As s_{ex}^2 / s_x^2 gets close to the unity, $b_1(ODR) / b_1(OLS)$ grows rapidly up to infinity, Figure 3.

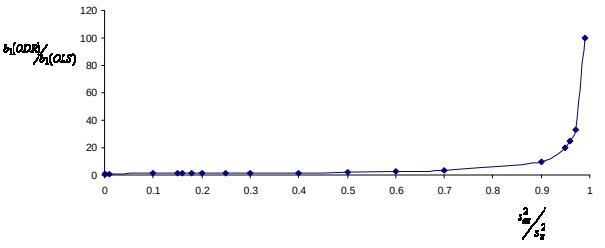

Fig. 3. $b_1(ODR) / b_1(OLS)$ **tending to infinity**

3. CASE STUDIES

Two case studies using ODR are discussed, in this work. At first, a catalytic fluorimetric method is compared with a photometric technique for the determination of the level of phytic acid in urine samples and secondly, the regression of an analytical curve for the determination of copper content in waters by Flame Atomic Absorption Spectrometry (FAAS). In these examples, it is considered equal uncertainties in both variables.

4. RESULTS AND DISCUSSION

4.1. Comparison of two methods

The level of phytic acid in urine samples was determined by a catalytic fluorimetric (CF) method, and the results are compared with those obtained using an established extraction photometric (EP) technique. The results, in mgL^{-1}, are means of triplicate measurements, Table 2.

Table 2. Comparison of CF versus EP [3]

EP	CF	EP	CF
1.98	1.87	0.13	0.14
2.31	2.20	3.15	3.20
3.29	3.15	2.72	2.70
3.56	3.42	2.31	2.43
1.23	1.10	1.92	1.78
1.57	1.41	1.56	1.53
2.05	1.84	0.94	0.84
0.66	0.68	2.27	2.21
0.31	0.27	3.17	3.10
2.82	2.80	2.36	2.34

ODR line: $\hat{y} = -0.056 + 0.996x$

$CI(b_0) = -0.056 \pm 0.063 \, (-0.119, 0.007)$

$CI(b_1) = 0.996 \pm 0.040 \, (0.955, 1.036)$

Based on equation (9), as the confidence intervals include the optimal parameters for the slope b_1 and intercept b_0, 1 and zero respectively, from the orthogonal regression of the data from calibration, showing equivalence in between methods.

If *t*-test was used, what is incorrect, the comparison of the methods would not be considered equivalent, because, to the 95% confidence level because the values of *t* calculated (3.59) is higher than the *t* critical (2.09).

4.2. Data calibration for the determination of copper content in waters by FAAS

Data regression from Table 3 shows that there is no difference between ODR coefficients and the OLS one, because s_{ex}^2 / s_x^2 is very close to zero. So, in this case, the uncertainty in *x*-axis can be negligible in the regression of the analytical curve.

Table 3. Analytical curve for the determination of copper content in waters by FAAS

Concentration, mg mL^{-1}	Absorbance		
0.10	0.0081	0.0079	0.0080
0.25	0.0206	0.0205	0.0202
0.50	0.0391	0.0394	0.0398
0.75	0.0596	0.0591	0.0590
1.00	0.0782	0.0790	0.0792

OLS line: $\hat{y} = 0.0004 + 0.0784x$

ODR line: $\hat{y} = 0.0004 + 0.0784x$

5. CONCLUSION

This work shows us to different ratios of s_{ex}^2 / s_x^2, the proportion that ODR moves away OLS.

There is need to evaluate the impact of uncertainty on *x*-axis before performing linear regression once the inadequate application of regression may lead to different conclusions.

REFERENCES

[1] J. Tellinghuisen, *"Least Squares in Calibration: Dealing with Uncertainty in x"*, The Analyst, pp 1961-1969, 2010.

[2] V. Synek, *"Calibration Lines Passing Through the Origin with Errors in Both Axes"*, Accreditation and Quality Assurance, pp 360-367, 2001.

[3] J. N. Miller, J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, 4.ed. Harrow, U.K.: Prentice Hall, 2000.

[4] D. L. Massart et al., *Handbook of Chemometrics and Qualimetrics*, Part A. Amsterdam, Elsevier, 1997.

[5] K. Danzer, M. Wagner; C. Fischbacher, *"Calibration by orthogonal and common least squares - theoretical and practical aspects"*, Fresenius' Journal of Analytical Chemistry, 352, pp 407-412, 1995.

[6] P.H. Garthwaite, I.T. Jolliffe, B. Jones, (1995) – *Statistical Inference*. Prentice Hall International (UK) Limited, 1995.

[7] P. J. Cornbleet, N. Gochman, *"Incorrect Least-Squares Regression Coefficient in Method-Comparison Analysis"*, Clinical Chemistry, pp 432-438, 1979.

[8] J. Wienold et al., *"Elemental Analysis of Copper and Magnesium Alloy Samples Using IR-Laser Ablation in Comparison with Spark and Glow Discharge Methods"*, Journal of Analytical Atomic Spectrometry, pp 1570-1574, 2009.

[9] J. Mandel, *The Statistical Analysis of Experimental Data*, Dover Publications, New York, 1964.