Chilean Journal of Statistics
Vol. 7, No. 1, April 2016, 3–16

Regression
Research Paper

Symmetrical linear calibration model to replicated data

Francisco William P. Marciano, Betsabé G. Blas Achic* and Francisco José A. Cysneiros

Department of Statistics, Federal University of Pernambuco, Recife, Brazil.
(Received: 2 October 2015 · Accepted in final form: 17 April 2016)

Abstract

Most calibration models in the literature assume that the errors are normally distributed. Unfortunately, when there are outliers in the data, these models often have very poor performance. In this work, we address this problem and propose a linear calibration model for replicate measurement by assuming that the error model follows a family of symmetric distributions, which includes lighter and heavier tails distributions than the normal, such as the Student-t, power exponential and logistic of type II distributions. In presence of outliers, these distributions are more stable than the normal distribution. Likelihood-based methodology is used to estimate the model parameters and the Fisher information matrix is used to construct confidence intervals for the unknown value of interest. Furthermore, a simulation study is carried out to verify the asymptotic properties of the proposed model main parameter. Finally, we present an application of the proposed model using a real data set from chemical analysis.

Keywords: atypical observation · chemical analysis · linear calibration model · symmetrical distribution.

Mathematics Subject Classification: Primary 62J05 · Secondary 62J99.

1. Introduction

The calibration models in analytical chemistry are generally used to determine the amount of an analyte in samples with unknown concentrations. In the chemical laboratory, a measurement process is performed in order to obtain the data set. Firstly, a series of samples is prepared with known concentrations of an analyte (standard samples) and also samples with an unknown concentration (test samples). The standard sample concentrations should cover, at least, the range of concentration encountered during the analysis of test samples and be evenly spaced across the range. In chemical analysis the sampling costs can be quite expensive, so it is usually encountered in the routine analysis a small number of samples. Analyzing each of these standards and test samples using a chosen technique, for instance, chromatography in case of organic chemistry analysis and atomic absorption spectroscopy for metals analysis, it will produce a series of measurements. For most analysis a plot of instrument response versus analyte concentration will show a linear relationship.

Calibration model for chemical analysis can be defined by the relationship between the instrument response and the analyte concentration. The data set for the first stage is

*Corresponding author. Email: betsabe@de.ufpe.br

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

obtained by recording the responses (Y) from standard concentrations (X), where these standards may be prepared in the laboratory or available from a commercial source. In the second stage, the response variables (Y_0) are recorded from the test sample with the unknown concentration (X_0). Deviations from linearity, however, are not uncommon, especially as the concentration of metallic analytes, for example, increases due to variance reasons, such as unabsorbed radiation, stray light, or disproportionate decomposition of molecules at high concentrations.

In order to have sufficient information related to the first stage from the calibration model, it is prepared a minimum of four standard samples. The estimates can be poor when the number of standard samples in the first stage from the calibration model is small, but the accuracy of the estimates increases if the number of standard samples increases. Usually, it is prepared two or three replicates of the standard samples. So, the analyst replicates measurements to ensure linearity, to improve the confidence in the result, to assess the variability in the analysis and to avoid a gross error in the analysis of a single aliquot. Response measurements of an analyte can be achieved by instrumental repetitions or by authentic repetitions, see Pimentel and Neto (1996). The instrumental repetitions are referred to repeated measurement responses on the same standard solution or analytical sample, whereas that the authentic repetitions are replicate measurements, which are carried out on different standard solutions for the same concentration level, so this last kind of experiment is more expensive. Therefore, replicates increases considerably the number of experimental measures and these information can be added in the calibration model as it will be discussed in the Section 2.

In statistical literature about linear calibration models usually it is assumed that the random errors follow the normal distribution. For example, Krutchkoff (1967) compares the classical and the inverse estimators of the normal calibration model by using Monte Carlo method. In Shukla (1972), they are compared by using the mean squared error (MSE), and also it is provided expressions for the bias, variance and MSE of the linear calibration model. In Bolfarine et al. (1996) and Blas et al. (2013) and Blas et al. (2007), it is studied the linear calibration model by taking into account measurement errors in the independent variable. Blas and Sandoval (2010) proposed a kind of generalization of the model discussed in Blas et al. (2007) by assuming heteroscedasticity. Nevertheless, it is known that atypical observations can have significant influence on inferences of statistical models with normally distributed errors. Thus, the main purpose of this paper is to consider the case when the random error have heavier tails than one normal distribution. A heavy tail distribution is one whose extreme probabilities approach zero relatively slowly. Therefore, we propose the linear calibration model for replicated data with the symmetric distributions family as assumption of the error models. Many distributions belong to this class such as the normal, Student-t, power exponential, logistics of type I and II, among others.

In the decade of 1970, several researchers began to develop statistical models with symmetric distributions inspired by the Kelker's work (Kelker (1970)). Theoretical and applied aspects of this class of distributions have been widely discussed in recent decades, see for example, Fang et al. (1990), Fang and Zhang (1990) and Fang and Anderson (1990).

1.1 Symmetric distribution

The general form of the classes of univariate symmetric probability density functions is defined as follows: the random variable Y has symmetric distribution with location parameter $\mu \in \mathrm{Re}$ and scale parameter $\phi > 0$, if the probability density function is given

by

$$f(y;\mu,\phi) = \frac{1}{\sqrt{\phi}}\, h\left[\frac{(y-\mu)^2}{\phi}\right], \quad y \in \mathrm{Re}, \tag{1}$$

for some positive function $h(\cdot)$ called generating density function,it is defined on Re^+ and $\int_0^{+\infty} u^{-\frac{1}{2}} h(u)\, du = 1$. This condition guarantees that $f(y;\mu,\phi)$ is a density function (see Fang et al. (1990)). We denote the density function given in (1) by $S(\mu,\phi,h)$.

The insertion point of this work is motivated by both aspects the inadequacy of the normal distribution in the presence of atypical observations and the necessity of replicate measurements. In this work, we discuss the Student-t, power exponential and logistic of type II (Log-II) distributions, which stem from the class of distributions defined in (1). These distributions are suggested as an alternative to the normal distribution.

There exist some works in literature that have studied heavy tailed distributions as an alternative to normal distribution for the errors from linear calibration models. For example, in Branco et al. (1998) it is considered the calibration model problem assuming Student-t errors under a Bayesian perspective. Branco et al. (2000) discussed the Bayesian calibration model with the assumption that the error distribution is elliptically symmetric. Lima et al. (2007) studied the calibration model with measurement error assuming Student-t errors. Figueiredo et al. (2008) presented the Bayesian calibration model assuming that the random errors follow a skew normal distribution. Figueiredo et al. (2010) introduced the EM algorithm to find the maximum likelihood estimates of a linear calibration model assuming that the errors follow the skew normal distribution, and Blas et al. (2013) studied the linear calibration model with Berkson type measurement errors assuming that the error follows the normal distribution and also considered replicate measurements. There are calibration model approaches that present less restrictive assumptions over error specifications. For instance, Lwin (1981) obtained approximated expressions for the MSEs of the calibration model estimators assuming finite fourth moment of the errors.

In this work, our focus are on considering replicate measurements on the linear calibration model with the symmetric error distributions as given in (1), and studying the asymptotic behavior of the estimator of X_0 for the Student-t, power exponential and Log-II distributions. The power exponential and Log-II distributions are not discussed in the works reported in the literature related to calibration models. On the other hand, by considering replicate measurements we are adding more information over the modeling, so our proposed model would be closer to the experiment reality.

The rest of the paper is organized as follows. In Section 2, we present the heavy tailed linear calibration model (hereafter called the "Proposed-M"), and discuss the parameter estimation. Section 3 presents the simulation study of the Proposed-M to evaluate the asymptotic behavior of X_0 under three distributions for the error model: Student-t, power exponential and Log-II. Furthermore, confidence intervals and coverage probabilities are obtained for the parameter X_0 in different scenarios evaluated. In Section 4, we present an application to verify the suitability of the Proposed-M. And finally, some concluding remarks are presented in Section 5.

2. Heavy tailed linear calibration model

The heavy tailed linear calibration model with replicate measurements (Proposed-M) is defined by both the first and second stages which are given by the following equations, respectively,

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}, \quad j = 1, \cdots, m_i \quad \text{and} \quad i = 1, \cdots, n \tag{2a}$$

$$Y_{0i} = \mu_0 + \epsilon_{0i}, \quad i = n+1, \cdots, n+r, \tag{2b}$$

where $\mu_{ij} = \beta_0 + \beta_1 X_i$ and $\mu_{0i} = \beta_0 + \beta_1 X_0$ are the systematic components from the first and second stages, respectively.

In the first stage (2a), the variables Y_{ij}, with $j = 1, \cdots, m_i$ and $i = 1, \cdots, n$ are observed. The quantities X_i are known fixed values and the random errors $\epsilon_{ij} \overset{iid}{\sim} S(0, \phi, h)$, where *iid* means independent and identically distributed and " $\sim$ " means follows. In the second stage (2b), we have the responses Y_{0i} from an instrument as a function of the unknown concentration X_0 of test samples and $\epsilon_{0i} \overset{iid}{\sim} S(0, \phi, h_0)$. The error variables ϵ_{ij} and ϵ_{0i} are mutually uncorrelated. The functions h and h_0 define the distribution belonging to the class of symmetric distributions. In practice, the assumption $h = h_0$ is made when Y_{ij} and Y_{0i} are obtained using the same method. If there are outliers in the first or second stage the assumption that $h \neq h_0$ can be suitable. The model parameters are β_0, β_1, X_0, ϕ and the main interest is to estimate X_0.

The density of Y_{ij} and Y_{0i} are given by

$$f_{Y_{ij}}(y_{ij}) = \frac{1}{\sqrt{\phi}}\, h(u_{ij}) \quad \text{and} \quad f_{Y_{0i}}(y_{0i}) = \frac{1}{\sqrt{\phi}}\, h_0(u_{0i}),$$

where $u_{ij} = (y_{ij} - \mu_{ij})^2/\phi$, $u_{0i} = (y_{0i} - \mu_{0i})^2/\phi$ with $Y_{ij} \sim S(\mu_{ij}, \phi, h)$ and $Y_{0i} \sim S(\mu_{0i}, \phi, h_0)$.

The log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, X_0, \phi)^\top$ is given by

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log\phi \left(\sum_{i=1}^{n} m_i + r\right) + \sum_{i=1}^{n}\sum_{j=1}^{m_i} \log h(u_{ij}) + \sum_{i=n+1}^{n+r} \log h_0(u_{0i}). \tag{3}$$

The score function is constructed as follows: assuming that the functions h and h_0 are continuous and differentiable we can define $W_h(u) = \partial \log[h(u)]/\partial u$. Thus, the score functions for the Proposed-M are given by

$$\mathbf{U}(\beta_0) = \tfrac{\partial l(\theta)}{\partial \beta_0} = -\frac{2}{\phi}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} W_h(u_{ij})(y_{ij} - \mu_{ij}) + \sum_{i=n+1}^{n+r} W_{h_0}(u_{0i})(y_{0i} - \mu_{0i})\right\}$$

$$\mathbf{U}(\beta_1) = \tfrac{\partial l(\theta)}{\partial \beta_1} = -\frac{2}{\phi}\left\{\sum_{i=1}^{n} x_i \sum_{j=1}^{m_i} W_h(u_{ij})(y_{ij} - \mu_{ij}) + X_0 \sum_{i=n+1}^{n+r} W_{h_0}(u_{0i})(y_{0i} - \mu_{0i})\right\}$$

$$\mathbf{U}(X_0) = \tfrac{\partial l(\theta)}{\partial X_0} = -\frac{2\beta_1}{\phi} \sum_{i=n+1}^{n+r} W_{h_0}(u_{0i})(y_{0i} - \mu_{0i})$$

$$\mathbf{U}(\phi) = \tfrac{\partial l(\theta)}{\partial \phi} = -\frac{1}{2\phi}\left[\sum_{i=1}^{n} m_i + r\right] - \frac{1}{\phi}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} W_h(u_{ij})u_{ij} + \sum_{i=n+1}^{n+r} W_{h_0}(u_{0i})u_{0i}\right\}.$$

We can find the expression $W_h(u)$ for the normal, Student-t, power exponential and Log-II symmetric distributions in Cysneiros and Paula (2005), which are presented in Table 1.

Table 1. Values of $W_h(u)$, d_h and f_h for the normal, Student-t, power exponential and Log-II distributions.

Distribution	$W_h(u)$	d_h	f_h
Normal	$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
Student-t (v)	$-\frac{(v+1)}{2(v+u)}$	$\frac{(v+1)}{4(v+3)}$	$\frac{3(v+1)}{4(v+3)}$
Power exponential (k)	$-\frac{1}{2(k+1)u^{k/(k+1)}}$	$\frac{\Gamma(\frac{3-k}{2})}{4(2^{k-1})(k+1)^2\Gamma(\frac{k+1}{2})}$	$\frac{(k+3)}{4(k+1)}$
Logistic-II	$-\frac{\exp(-\sqrt{u})-1}{(-2\sqrt{u})[1+\exp(-\sqrt{u})]}$	$\frac{1}{12}$	0.60749

The maximum likelihood estimators of the Proposed-M can not be obtained in analytical form, so it is necessary some numerical maximization such as Newton-Raphson or Fisher scoring method (Nocedal and Wright (1999)). In the following we will describe an alternative form to find the maximum likelihood estimates from the Proposed-M.

Let $\boldsymbol{Z}$ be an $(\sum_{i=1}^{n} m_i + r) \times 2$ matrix, with the first column vector of ones and the second column containing the vector $(X_1\mathbf{1}_{m_1}^\top, X_2\mathbf{1}_{m_2}^\top, \cdots, X_n\mathbf{1}_{m_n}^\top, X_0\mathbf{1}_r^\top)^\top$, where $\mathbf{1}_a$ define an a-dimensional column vector of ones. It is defined the matrix $\boldsymbol{W} = (\boldsymbol{Y}^\top, \boldsymbol{Y_0}^\top)^\top$, which is a vector containing the response variables from the model (2a-2b), where $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \cdots, \mathbf{Y}_n^\top)^\top$ with $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{im_i})^\top$, and $\mathbf{Y}_0 = (Y_{0n+1}, Y_{0n+2}, \cdots, Y_{0n+r})^\top$ are the vector of observed responses from the first and second stages, respectively. So, we can re-writte the model (2a-2b) as

$$\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_{im_i}, \cdots, \epsilon_{im_n}, \epsilon_{0n+1}, \cdots, \epsilon_{0n+r})^\top$. It must be noted that the parameter X_0 is in $\boldsymbol{Z}$, so an iterative procedure is therefore proposed to obtain the maximum likelihood estimators of the Proposed-M which is given by the following iterative procedure,

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{Z}^{(k)\top}\mathbf{D}(v^{(k)})\mathbf{Z}^{(k)}\}^{-1}\mathbf{Z}^{(k)\top}\mathbf{D}(v^{(k)})\mathbf{y}, \tag{5}$$

$$\phi^{(k+1)} = \frac{1}{N+r}\{\mathbf{y} - \mathbf{Z}^{(k)}\boldsymbol{\beta}^{(k)}\}^\top\mathbf{D}(v^{(k)})\{\mathbf{y} - \mathbf{Z}^{(k)}\beta^{(k)}\}, \tag{6}$$

$$X_0{}^{(k+1)} = \text{argmax}_{X_0}\{l(\boldsymbol{\beta}^{(k+1)}, \phi^{(k+1)}, X_0^{(k)})\}. \tag{7}$$

where $N = \sum_{i=1}^{n} m_i$ and $\mathbf{D}(v^{(k)}) = \text{diag}\{v_1^{(k)}, v_2^{(k)}, \cdots, v_n^{(k)}, v_{n+1}^{(k)}, \cdots, v_{n+r}^{(k)}\}$ with $v_i^{(k)} = -2W_h^{(k)}(\cdot)$ for $i = 1, \cdots, n$ and $v_i^{(k)} = -2W_{h_0}^{(k)}(\cdot)$ for $i = n+1, \cdots, n+r$ and $k = 0, 1, \ldots$. Initial values for iterative procedure are given by the maximum likelihood estimates for the usual calibration model (Shukla (1972)).

Since in the equations (5) and (6) the matrices $\mathbf{Z}$ and $\mathbf{D}$ depend only on the value X_0, we can substitute them into the logarithm of the likelihood function (3) written in matrix form using (4), such that it will only depend upon the parameter X_0. Hence, it can be maximized using some numerical method as given in (7), where the starting value of X_0 can be used from the usual calibration model (Shukla (1972)) and the maximizing method BFGS can be used, which can be implemented in the software R. Once the parameter X_0 is estimated from Equation (7), in the next iteration step the estimates of β and ϕ can be found by using this estimate in the equations (5) and (6), respectively.

The Fisher information matrix $\boldsymbol{K_\theta}$ for the Proposed-M can be expressed as follows

$$\boldsymbol{K}_{\boldsymbol{\theta}} = \mathbb{E}\left[\left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right] = \begin{bmatrix} K_{\beta_0} & K_{\beta_0\beta_1} & K_{\beta_0 X_0} & K_{\beta_0\phi} \\ & K_{\beta_1} & K_{\beta_1 X_0} & K_{\beta_1\phi} \\ & & K_{X_0} & K_{X_0\phi} \\ & & & K_{\phi} \end{bmatrix},$$

with the matrix components given by

$$K_{\beta_0} = \frac{4}{\phi}\left\{N d_h + r d_{h_0}\right\}, \qquad K_{\beta_0\beta_1} = \frac{4}{\phi}\left\{\sum_{i=1}^{n} m_i X_i d_h + r X_0 d_{h_0}\right\},$$

$$K_{\beta_1} = \frac{4}{\phi}\left\{\sum_{i=1}^{n} m_i X_i^2 d_h + r X_0^2 d_{h_0}\right\}, \qquad K_{\beta_0 X_0} = \frac{4}{\phi} r\beta_1 d_{h_0},$$

$$K_{X_0} = \frac{4}{\phi} r\beta_1^2 d_{h_0}, \qquad K_{\beta_1 X_0} = \frac{4}{\phi} r\beta_1 X_0 d_{h_0},$$

$$K_{\phi} = \frac{1}{4\phi^2}\left\{N(4f_h - 1) + r(4f_{h_0} - 1)\right\}, \qquad K_{\beta_0\phi} = K_{\beta_1\phi} = K_{X_0\phi} = 0,$$

where the values $d_h = \mathbb{E}[W_h^2(U)U]$, $d_{h_0} = \mathbb{E}[W_{h_0}^2(U_0)U_0]$, $f_h = \mathbb{E}[W_h^2(U)U^2]$ and $f_{h_0} = \mathbb{E}[W_{h_0}^2(U_0)U_0^2]$ are computed for some distribution as given in Table 1.

Under conditions that are fulfilled for parameters in the interior of the parameter space but not on the boundary, the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is normal with zero mean and the variance is the inverse of the expected information matrix. Then, in order to construct a confidence interval (CI) for X_0 we consider

$$\frac{\widehat{X}_0 - X_0}{\sqrt{\mathrm{Var}(\widehat{X}_0)}} \xrightarrow{D} N(0,1),$$

where $\mathrm{Var}(\widehat{X}_0)$ is the variance of $\widehat{X}_0$ derived from the Fisher information matrix. Therefore, the approximate CI for X_0 with confidence coefficient $(1-\alpha)$ is given by

$$\left[\widehat{X}_0 - z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{X}_0)},\ \widehat{X}_0 + z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{X}_0)}\right],$$

where $z_{\frac{\alpha}{2}}$ is the quantile of order $(1-\frac{\alpha}{2})$ of the standard normal distribution.

3. Simulation study

The behavior of the maximum likelihood estimator of X_0 was studied through the empirical bias and the mean square error (MSE), for which it is generated 10000 Monte Carlo samples from the Proposed-M with sample sizes $n = 5$, 20, 40 and 100 for the first stage and $r = 3$, 20 and 100 for the second stage. The values of the parameters were $\beta_0 = 0.1$, $\beta_1 = 2$ and $\phi = 0.04$. The values x_i, $i = 1, \cdots, n$, are taken to be equally spaced values over the range $[0, 2]$, being the values $x_1 = 0$ and $x_i = x_{i-1} + 2/(n-1)$ for $i = 2, \ldots, n$. The parameter

values X_0 were also chosen over the range $[0, 2]$ such that 0.01 (extreme inferior value), 0.8 (near to the central value) and 1.9 (extreme superior value). The empirical mean bias and the MSE are given by $\sum_{i=1}^{10000}(\hat{X}_0 - X_0)/10000$ and $\sum_{i=1}^{10000}(\hat{X}_0 - X_0)^2/10000$, respectively. The simulation results were obtained using the *software* R version 2.11.1.

Tables 2-5 present both the empirical bias and the MSE of X_0 for the normal, Student-t, power exponential and logistic-II distributions. These tables show that for all r and X_0, the bias and the MSE decrease as the size of n increases, and when X_0 is close to the midpoint of the interval $[0, 2]$ they are much smaller. In Tables 2-4 , we observe that for all n and X_0 , the MSE decreases as the size of r increases, and for the logistic-II distribution this behavior remains for most of the cases, except for the smaller sample size of n and the extremes values of X_0.

Table 2. Empirical bias and MSE of $\hat{X}_0$ for the normal distribution.

		$r = 3$		$r = 20$		$r = 100$	
X_0	n	Bias	MSE	Bias	MSE	Bias	MSE
0.01	5	-0.0034	0.0093	-0.0028	0.0063	-0.0054	0.0060
	20	-0.0012	0.0052	-0.0009	0.0023	-0.0014	0.0019
	40	-0.0014	0.0043	-0.0006	0.0015	-0.0005	0.0010
	100	-0.0001	0.0037	0.0002	0.0009	-0.0003	0.0005
0.8	5	-0.0012	0.0055	-0.0002	0.0026	-0.0007	0.0021
	20	0.0012	0.0039	-0.0003	0.0010	-0.0006	0.0006
	40	0.0003	0.0036	-0.0001	0.0008	0.0000	0.0004
	100	0.0001	0.0035	-0.0001	0.0006	-0.0001	0.0002
1.9	5	0.0035	0.0086	0.0038	0.0057	0.0025	0.0043
	20	0.0014	0.0049	0.0009	0.0021	0.0012	0.0015
	40	-0.0004	0.0042	0.0005	0.0013	0.0006	0.0008
	100	-0.0005	0.0037	0.0001	0.0008	0.0004	0.0004

Table 3. Empirical bias and MSE of $\hat{X}_0$ for the Student-t distribution.

		$r = 3$		$r = 20$		$r = 100$	
X_0	n	Bias	MSE	Bias	MSE	Bias	MSE
0.01	5	-0.0161	0.0384	-0.0152	0.0266	-0.0089	0.0219
	20	-0.0027	0.0170	-0.0039	0.0070	-0.0039	0.0059
	40	-0.0005	0.0143	-0.0013	0.0042	-0.0016	0.0031
	100	-0.0014	0.0121	-0.0004	0.0026	-0.0009	0.0014
0.8	5	-0.0014	0.0208	-0.0036	0.0102	-0.0017	0.0081
	20	-0.0020	0.0128	-0.0002	0.0031	-0.0005	0.0019
	40	0.0001	0.0120	-0.0002	0.0022	-0.0001	0.0011
	100	-0.0017	0.0114	0.0002	0.0018	0.0000	0.0006
1.9	5	0.0150	0.0344	0.0125	0.0227	0.0097	0.0178
	20	0.0046	0.0160	0.0038	0.0062	0.0043	0.0047
	40	0.0047	0.0135	0.0014	0.0038	0.0023	0.0026
	100	0.0007	0.0124	0.0012	0.0024	0.0005	0.0012

Table 4. Empirical bias and MSE of $\hat{X}_0$ for the power exponential distribution.

		$r=3$		$r=20$		$r=100$	
X_0	n	Bias	MSE	Bias	MSE	Bias	MSE
0.01	5	-0.0308	0.0866	-0.0316	0.0609	-0.0213	0.0484
	20	-0.0100	0.0378	-0.0091	0.0154	-0.0074	0.0126
	40	-0.0043	0.0303	-0.0066	0.0091	-0.0034	0.0066
	100	-0.0025	0.0259	-0.0029	0.0053	-0.0018	0.0029
0.8	5	-0.0030	0.0431	-0.0073	0.0234	-0.0072	0.0179
	20	-0.0021	0.0283	-0.0016	0.0066	-0.0027	0.0041
	40	-0.0027	0.0258	-0.0018	0.0048	-0.0011	0.0023
	100	-0.0023	0.0237	0.0001	0.0037	-0.0001	0.0012
1.9	5	0.0266	0.0713	0.0309	0.0514	0.0262	0.0439
	20	0.0022	0.0349	0.0072	0.0139	0.0078	0.0103
	40	0.0029	0.0289	0.0028	0.0082	0.0035	0.0056
	100	-0.0005	0.0263	0.0017	0.0050	0.0016	0.0025

Table 5. Empirical bias and MSE of $\hat{X}_0$ for the logistic-II distribution.

		$r=3$		$r=20$		$r=100$	
X_0	n	Bias	MSE	Bias	MSE	Bias	MSE
0.01	5	-0.0586	0.4860	-0.0513	0.1014	-0.0427	0.0832
	20	-0.0140	0.0570	-0.0152	0.0262	-0.0155	0.0226
	40	-0.0019	0.0467	-0.0067	0.0150	-0.0061	0.0111
	100	-0.0024	0.0391	-0.0017	0.0089	-0.0025	0.0049
0.8	5	-0.0120	0.0724	-0.0112	0.0341	-0.0074	0.0289
	20	-0.0048	0.0436	-0.0021	0.0113	-0.0024	0.0067
	40	0.0017	0.0386	-0.0032	0.0081	-0.0021	0.0038
	100	-0.0005	0.0362	-0.0008	0.0061	-0.0003	0.0021
1.9	5	0.0503	0.2543	0.0469	0.0995	0.0387	0.0774
	20	0.0135	0.0543	0.0125	0.0239	0.0127	0.0183
	40	0.0065	0.0444	0.0082	0.0138	0.0053	0.0094
	100	0.0029	0.0380	0.0045	0.0084	0.0021	0.0044

Tables 6-7 present the lower limit (LL) and upper limit (UL) of the asymptotic 95% confidence interval for the parameter X_0, and the covering percentage for the normal, Student-t, power exponential and logistic-II distributions, respectively. Analyzing these tables, we observe that for all r, when $X_0 = 0.01$ the LL and UL decrease as the size of n increases, and when $X_0 = 1.9$ and for, in most of the cases, $X_0 = 0.8$ the LL increases and UL decreases as the size of n increases. This causes the covering percentage to increase approaching 95%.

Table 6. 95% asymptotic confidence interval and coverage probability for the parameter X_0 under the normal and Student-t distributions.

Normal													
		LL				UL				Coverage			
X_0	r	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100
0.01	3	0.0055	0.0022	0.0014	0.0010	0.1495	0.1394	0.1312	0.1277	0.827	0.915	0.935	0.944
0.8	3	0.6894	0.6877	0.6862	0.6874	0.9089	0.9127	0.9125	0.9137	0.819	0.917	0.935	0.945
1.9	3	1.7641	1.7744	1.7798	1.7839	2.0399	2.0289	2.0221	2.0173	0.823	0.919	0.936	0.944
0.01	20	0.0025	0.0013	0.0009	0.0008	0.1541	0.0998	0.0822	0.0669	0.921	0.933	0.942	0.943
0.8	20	0.7047	0.7388	0.7471	0.7523	0.8935	0.8606	0.8535	0.8478	0.919	0.938	0.940	0.946
1.9	20	1.7639	1.8147	1.8309	1.8449	2.0418	1.9871	1.9697	1.9567	0.924	0.934	0.942	0.946
0.01	100	0.0016	0.0011	0.0009	0.0007	0.1559	0.0937	0.0727	0.0524	0.947	0.944	0.946	0.950
0.8	100	0.7071	0.7504	0.7620	0.7717	0.8915	0.8494	0.8374	0.8283	0.951	0.954	0.949	0.950
1.9	100	1.7606	1.8214	1.8422	1.8598	2.0445	1.9813	1.9603	1.9412	0.960	0.954	0.952	0.948
Student-t													
		LL				UL				Coverage			
X_0	r	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100
0.01	3	0.0122	0.0041	0.0032	0.0029	0.2434	0.2254	0.2144	0.2045	0.784	0.902	0.920	0.922
0.8	3	0.6101	0.6093	0.6115	0.6116	0.9859	0.9898	0.9917	0.9905	0.782	0.901	0.919	0.924
1.9	3	1.6764	1.6873	1.6988	1.7079	2.1500	2.1158	2.1066	2.0996	0.780	0.908	0.916	0.927
0.01	20	0.0059	0.0024	0.0015	0.0011	0.2485	0.1599	0.1289	0.1064	0.903	0.922	0.938	0.944
0.8	20	0.6368	0.6966	0.7099	0.7199	0.9573	0.9020	0.8886	0.8801	0.905	0.928	0.934	0.943
1.9	20	1.6765	1.7580	1.7841	1.8068	2.1576	2.0485	2.0174	1.9947	0.902	0.930	0.937	0.944
0.01	100	0.0050	0.0019	0.0013	0.0008	0.2570	0.1495	0.1141	0.0804	0.926	0.937	0.940	0.948
0.8	100	0.6407	0.7149	0.7368	0.7525	0.9543	0.8816	0.8634	0.8473	0.923	0.939	0.942	0.948
1.9	100	1.6666	1.7685	1.8028	1.8330	2.1508	2.0380	2.0012	1.9693	0.934	0.946	0.951	0.946

Table 7. 95% asymptotic confidence interval and coverage probability for the parameter X_0 under the power exponential and logistic-II distributions.

power exponential													
		LL				UL				Coverage			
X_0	r	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100
0.01	3	0.0166	0.0068	0.0051	0.0043	0.3449	0.3141	0.2997	0.2868	0.794	0.895	0.906	0.911
0.8	3	0.5305	0.5271	0.5275	0.5291	1.0709	1.0706	1.0678	1.0659	0.794	0.888	0.902	0.912
1.9	3	1.5764	1.5967	1.6129	1.6213	2.2762	2.2083	2.1927	2.1774	0.804	0.895	0.905	0.908
0.01	20	0.0089	0.0031	0.0024	0.0014	0.3545	0.2186	0.1805	0.1448	0.903	0.926	0.929	0.938
0.8	20	0.5631	0.6513	0.6718	0.6865	1.0263	0.9445	0.9256	0.9138	0.909	0.925	0.934	0.931
1.9	20	1.5749	1.7009	1.7358	1.7681	2.2795	2.1171	2.0677	2.0337	0.903	0.923	0.930	0.941
0.01	100	0.0074	0.0029	0.0020	0.0012	0.3634	0.2062	0.1563	0.1100	0.926	0.934	0.936	0.942
0.8	100	0.5658	0.6795	0.7089	0.7325	1.0210	0.9167	0.8886	0.8669	0.931	0.937	0.944	0.945
1.9	100	1.5658	1.7155	1.7615	1.8037	2.2837	2.1015	2.0440	1.9971	0.934	0.938	0.940	0.945

logistic-II													
		LL				UL				Coverage			
X_0	r	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100	n=5	n=20	n=40	n=100
0.01	3	0.0190	0.0056	0.0042	0.0035	0.5265	0.4185	0.3940	0.3769	0.828	0.922	0.933	0.936
0.8	3	0.4485	0.4380	0.4422	0.4422	1.3010	1.1587	1.1578	1.1546	0.833	0.916	0.926	0.935
1.9	3	1.4543	1.5054	1.5207	1.5349	2.4570	2.3251	2.2940	2.2724	0.833	0.911	0.930	0.937
0.01	20	0.0092	0.0040	0.0024	0.0018	0.4848	0.2909	0.2330	0.1902	0.922	0.932	0.942	0.942
0.8	20	0.4777	0.6007	0.6310	0.6483	1.1139	0.9926	0.9695	0.9500	0.934	0.938	0.944	0.944
1.9	20	1.4528	1.6349	1.6839	1.7253	2.4365	2.1945	2.1273	2.0787	0.919	0.935	0.942	0.944
0.01	100	0.0076	0.0033	0.0022	0.0014	0.4952	0.2718	0.2047	0.1418	0.937	0.943	0.946	0.944
0.8	100	0.4883	0.6382	0.6780	0.7096	1.1077	0.9563	0.9178	0.8883	0.956	0.947	0.952	0.953
1.9	100	1.4430	1.6529	1.7173	1.7739	2.4377	2.1698	2.0938	2.0314	0.939	0.947	0.948	0.949

4. Application

We apply the Proposed-M to a data set from Neto et al. (2007). In this application the main interest is to estimate the zinc concentration (X_0). This data set is presented in Table 8 and it consists on aqueous solutions containing five levels of zinc concentration (X) and their related replicate measurements (Y) of size $m_i = 3$, which are obtained from the flame atomic absorption spectrometry analytical procedure.

Table 8. Zinc concentration levels and absorbance.

Concentration	Replicate measurements		
X	Y_1	Y_2	Y_3
0.0	0.696	0.696	0.706
0.5	7.632	7.688	7.603
1.0	14.804	14.861	14.731
2.0	28.895	29.156	29.322
3.0	43.993	43.574	44.699

To show the suitability of our approach in a practical experimentation, we consider the zinc standard concentration value 1.0 to be an unknown value X_0, and then the related response variable Y_0 will be considered the second stage data on the calibration model. So, the rest of the data set are considered as the first stage data on the calibration model. The data set was obtained using only an analytical method, and then it belongs to the same population and it is reasonable to consider $h = h_0$ in (3), i.e., the dependent variable from the first and second stage have the same distribution.

The Proposed-M defined in Section 2 is fitted to this data set over the normal, Student-t, power exponential and logistic-II distributions. The values of the parameters v and k related to the Student-t and power exponential distributions, respectively, were chosen in a selection procedure based on the Akaike information criterion (AIC). Based on this selection procedure, the values of the parameters chosen were $v = 3$ for the Student-t distribution and $k = 0.5$ for the power exponential distribution.

Table 9 presents the Akaike information criterion (AIC), Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQ) which are model selection criteria. Based on these criteria, the selected model is the one with minimum value. We observe that the values of the AIC, BIC and HQ criteria are smaller for the fitted Proposed-M assuming Student-t errors as compared with those values of the other fitted models.

Table 9. AIC, BIC and HQ criteria for the Proposed-M using the zinc data set from Neto et al. (2007).

Distribution	AIC	BIC	HQ
Normal	6.8603	8.2764	6.8452
Student-t ($v = 3$)	3.8399	5.2560	3.8249
Power exponential ($k = 0.5$)	5.4678	6.8839	5.4528
logistic-II	5.5385	6.9546	5.5234

Table 10. Proposed-M parameters estimates, standard error and confidence interval amplitude $U(X_0)$ under errors normal, Student-t, power exponential and logistic-II .

Distribution	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{X}_0$	$\hat{\phi}$	$U(X_0)$
Normal	0.5159	14.4526	0.9882	0.0857	0.0259
	(0.1290)	(0.0708)	(0.0132)	(0.0313)	-
Student-t ($v = 3$)	0.5891	14.3357	0.9912	0.0204	0.0151
	(0.0746)	(0.0410)	(0.0077)	(0.0099)	-
Power exponential ($k = 0.5$)	0.5540	14.3845	0.9906	0.0283	0.0220
	(0.1088)	(0.0598)	(0.0112)	(0.0127)	-
logistic-II	0.5539	14.3929	0.9897	0.0199	0.0217
	(0.1076)	(0.0591)	(0.0110)	(0.0085)	-

The generated envelopes for the normal, Student-t, power exponential and logistic-II distributions, as proposed by Atkinson (1981), are presented in Fig. 1. In this figure we observe that the Proposed-M with Student-t distribution fitted better the data set than the other distributions.

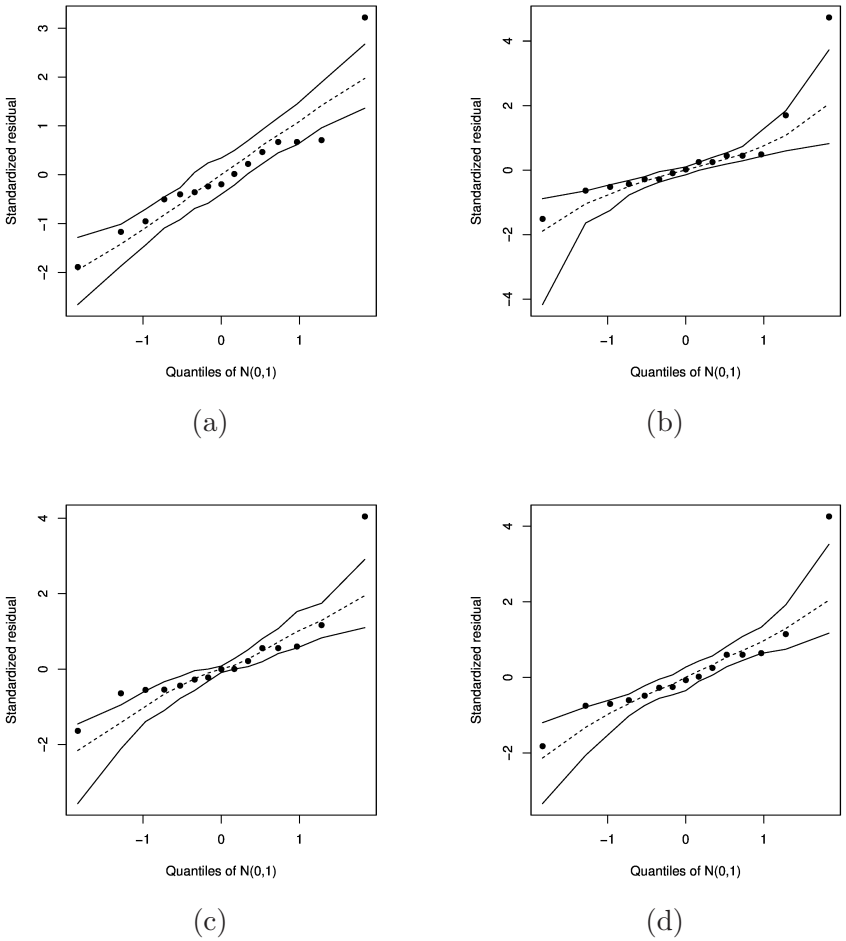

Figure 1. Simulated envelopes for the fitted Proposed-M using the zinc data set under (a) normal, (b) Student-t ($v = 3$), (c) power exponential ($k = 0.5$) and (d) logistic-II distributions.

5. Concluding remarks

A new calibration model with replicate measurements assuming a class of symmetrical distributions for the error model was proposed in this work. This new model is quite flexible to analyze data considering a class of symmetrical distributions instead of only the normal distribution. For example, error normality is very used in chemical analysis, but there are applications where normal distributions is not adequate. In this work, the simulation study was conducted to show the asymptotic behavior of the parameter of interest X_0 assuming Student-t, power exponential and logistic-II errors for the Proposed-M. It was observed that when the sample size of the first and second stage increase both the empirical bias and the MSE decrease for the three distributions. We also observed that for small sample sizes in the first stage ($n = 5$) for all the studied distributions, the estimator of X_0 has large bias when it is close to zero, and the lower value of bias occurs when the value of X_0 is near to the midpoint of the range of the variable X. The application example proved that the three alternative models of the normal model can be more appropriate for the data set according to the AIC, BIC and HQ criteria. Moreover, based on simulated envelopes, the Student-t model is more appropriate than the other two alternative models for the zinc data set.

Acknowledgements

The authors gratefully acknowledge financial support from the Brazilian Agency CNPq [201192/2015-2 to B.G.B.A.]. The authors are also very grateful to the referee and the associate editor for helpful comments and suggestions which resulted in this improved version of the paper.

References

Atkinson, A.C., 1981. Two graphical display for outlying and influential observations in regression. Biometrika, 68, 13–20.

Blas, B. G., Bolfarine, H., Lachos, V.H., 2013. Statistical analysis of controlled calibration model with replicates. Journal of Statistical Computation and Simulation, 83, 941–961.

Blas, B.G., Sandoval, M.C., Yoshida, O.S., 2007. Homoscedastic controlled calibration model. Journal of Chemometrics, 21, 145–155.

Blas, B.G., Sandoval, M.C., 2010. Heteroscedastic controlled calibration model applied to analytical chemistry. Journal of Chemometrics, 24, 241–248.

Bolfarine, H., Lima, C.R.O.P., Sandoval, M.C., 1997. Linear Calibration In Functional Regression Models. Communications in Statistics: Theory and Methods, 26, 2307–2328.

Branco, M.D., Bolfarine, H., Iglesias, P., 1998. Bayesian calibration under a Student-t model. Computational Statistics, 13, 319–338.

Branco, M.D., Bolfarine, H., Iglesias, P., Arellano-Valle, R.B., 2000. Bayesian analysis of the calibration problem under elliptical distributions. Journal of Statistical Planning and Inference, 90, 69–85.

Cysneiros, F.J.A., Paula, G.A., 2005. Restricted methods in symmetrical linear regression models. Computatinal Statistics & Data Analysis, 49, 689–708.

Fang, K.T., Kotz, S., Ng, K.W., 1990. Symmetric Multivariate and Related Distributions. Chapman and Hall, London.

Fang, K.T., Zhang, Y.T., 1990. Generalized Multivariate Analysis. Springer-Verlag, New York.

Fang, K.T., Anderson, T.W., 1990. Statistical Inference in Elliptical Contoured and Related Distributions. Allerton Press, New York.

Figueiredo, C.C., Sandoval, M.C., Bolfarine, H., Lima, C.R.O.P., 2008. Skew-normal linear calibration: a Bayesian perspective. Journal of Chemometrics, 22, 472–480.

Figueiredo, C.C., Bolfarine, H., Sandoval, M.C., Lima, C.R.O.P., 2010. On the skew-normal calibration model. Journal of Applied Statistics, 37, 435–451.

Kelker, D., 1970. Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhya A, 32, 419–430.

Krutchkoff, R., 1967. Classical and inverse regression methods of calibration. Technometrics, 9, 425–439.

Lange, K.L., Little, R., Taylor, J., 1989. Robust statistical modeling using t distribution. Journal of the American Statistical Association, 84, 881–896.

Lima, C.R.O.P., Sandoval, M.C., Bolfarine, H., Sousa, S.O., 2007. Robust estimation in calibration models using the student-t distribution. Journal of Applied Statistical Science, 15, 253–267.

Lwin, T., 1981. Discussion of hunter and lamboy's 1981 paper. Technometrics, 23, 339–341.

Neto, B.B., Scarminio, I.S., Bruns, R.E., 2007. Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria. Editora da Unicamp, Campinas.

Nocedal, J., Wright, S.J., 1999. Numerical optimization. Springer-Verlag, Nova York.

Pimentel, M.F., Neto, B.B., 1996. Calibração: uma revisão para químicos analíticos. Química Nova, 19, 268–277.

Shukla, G.K., 1972. On the problem of calibration. Technometrics, 14, 547–553.