http://dx.doi.org/10.5935/0103-5053.20130187

*J. Braz. Chem. Soc.*, Vol. 24, No. 9, 1469-1477, 2013.
Printed in Brazil - ©2013 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

# Construction of Analytical Curve Fit Models for Simvastatin using Ordinary and Weighted Least Squares Methods

***Flávia D. Marques-Marinho,*,[a] Ilka A. Reis[b] and Cristina D. Vianna-Soares[a]***

*[a]Department of Pharmaceutical Products, Faculty of Pharmacy and [b]Department of Statistics,
Institute of Exact Sciences, Federal University of Minas Gerais, Av. Pres. Antônio Carlos, 6627,
31270-901 Belo Horizonte-MG, Brazil*

Métodos analíticos requerem modelos adequados de ajuste de curva para expressar confiabilidade. Métodos dos mínimos quadrados ordinários ou ponderado (OLSM ou WLSM, respectivamente) foram usados para determinar o modelo matemático mais adequado à curva analítica, iniciando-se do método mais simples (linear) até o quadrático. A normalidade e a homocedasticidade dos resíduos dos modelos foram avaliadas. Curvas analíticas foram construídas pela injeção de 1, 5, 10, 15 e 20 µL de sinvastatina 40 µg mL$^{-1}$ (40, 200, 400, 600 e 800 ng) ou de 10 µL de sinvastatina 4, 20, 40, 60 e 80 µg mL$^{-1}$, empregando cromatografia líquida de alta eficiência com detecção por arranjo de diodos (λ 238 nm). Os melhores modelos foram o linear e o quadrático observados para os conjuntos de dados massas e concentrações, respectivamente. Na faixa de trabalho considerada, WLSM mostrou-se mais apropriado que OLSM. Os diferentes comportamentos apontam para a necessidade de uma escolha sensata do modelo mais adequado para expressar a curva analítica e assegurar a confiabilidade do método utilizado.

Analytical methods require adequate curve adjusting models to express reliability. The ordinary or weighted least squares methods (OLSM or WLSM, respectively) were used to determine the most adequate mathematical model for the analytical curve, beginning from the simplest method (linear) to quadratic. These models were evaluated with respect to normality and homoscedasticity of the residues. Analytical curves were built by injection of 1, 5, 10, 15 and 20 µL of simvastatin at 40 µg mL$^{-1}$ (40, 200, 400, 600 and 800 ng) or of 10 µL of simvastatin at 4, 20, 40, 60 and 80 µg mL$^{-1}$, employing high-performance liquid chromatography with photo-diode array detection (λ 238 nm). The best-adjusted models were the linear and the quadratic, when observed in terms of the dataset masses and concentrations, respectively. In the considered range, WLSM was more appropriate than OLSM. The different behavior indicated the need for careful selection of the most adequate model to express the analytical curve and the need for assuring the reliability of the used method.

**Keywords:** analytical curves, HPLC-UV, least squares method, weighted models, validation

## Introduction

Analytical methods require prior validation regardless of their field of application in order to obtain reliable results.[1] As an attempt to harmonize procedures, guidelines for analytical validation have been established by international or national regulatory and accreditation agencies, such as ANVISA and INMETRO in Brazil.[1-9] Pharmaceutical companies must comply with these requirements to ensure the efficacy, safety and quality of drugs.[10]

The parameters to be evaluated during a method validation depend on the assay purpose.[2,6,7,11] The commonly tested parameters are selectivity/specificity, accuracy, precision, response function (analytical curve) or linearity, range, limits of quantitation and of detection, and robustness.[12–15] The analytical curve expresses the relationship between the concentrations of the analyte and the detected responses within a range as a monotonic mathematical function (linear or non-linear).[16] Linearity is the ability of the analytical procedure to obtain results directly proportional (or by means of well-defined mathematical transformations) to the concentrations of the analyte in a specified range.[2,6,14] Until recently, the

*e-mail: flaviadmar@hotmail.com

analytical curve and linearity terms have been used in a confusing way because the function that expresses the response depends on the method.[12,15,17] However, when these methods cover a wide dynamical range, more complex or weighted models (quadratic, logarithmic, etc.) may be required.[12,16]

Regardless of the curve behavior, the linear regression obtained by the ordinary least squares method (OLSM) is the statistical method most applied to analytical procedures.[18] Nevertheless, OLSM has been indiscriminately used without evaluating the model and the assumptions related to its residuals.[19] OLSM requires the treatment of the outliers by using, for instance, the Jacknife test; as well as the verification of the assumptions of normality, homoscedasticity and independency of the residuals by the Ryan-Joiner or Jarque-Berra tests, the Levene test as modified by Brown-Forsy or Cook-Weisberg, and the Durbin-Watson test, respectively.[16,18,20] At least, two hypotheses should be satisfied: the normality of the response at every concentration level and the homogeneity of the variances of the responses (homoscedasticity) in the interval of the concentrations.[16] When a linear model (non-weighted) is not adequate for the selected range, the function of the analytical curve should be adjusted by testing mathematical models using either transformations in the responses or the weighted least squares method (WLSM).[12,13] WLSM is recommended when non-homoscedastic data are found, for instance, in a wide analytical range.[13,14] Regardless of the least squares method, a minimum of five concentration levels, in triplicate, is recommended to determine the linearity function since the uncertainty varies with the number of replicates.[7,15]

Alternative approaches to the OLSM-using weights ($1/x$, $1/x^2$ and $1/y$) have been described in the general validation of bioanalytical methods,[17,21,22] and applied to high-performance liquid chromatography (HPLC) coupled to mass spectrometer (LC-MS) for simvastatin (SIM) quantitation in human plasm.[23-26] To date, the linear regression obtained by OLSM has been reported for the validation of analytical methods, such as in HPLC-UV used to quantify SIM in bulk[27-30] or associated with other drugs.[31-34] The adequacy of the equation that represents the analytical curve is the most important way to assure low uncertainty in UV analytical measurements.[35]

Thus, in this work, different strategies have been employed to define the best analytical curve for SIM quantitation. Beginning from the simplest linear model that uses OLSM to more complex models that use WLSM, different injection volumes[36-38] or different concentrations were evaluated by HPLC-UV.

## Experimental

### Materials

The United States Pharmacopeia (USP) simvastatin reference standard (SIM RS, lot I0D382, 99.4% purity label claim, US Pharmacopeia, Rockville, MD, USA), phosphoric acid 85% (Merck, Darmstadt, Germany), methanol HPLC grade (Tedia, Fairfield, OH, USA), ultrapurified water (Direct Q3, Millipore, Bedford, MA, USA) and 0.45 µm filter membranes (Minisart RC15, Sartorius, Goettingen, Germany) were used.

### Instrumentation

An HP1200 quaternary liquid chromatography system equipped with automatic injector, column oven and ultraviolet diode array detector (UV/DAD), and ChemStation software version Rev.B.02.01-SR1 for data acquisition (Agilent, Palo Alto, CA, USA) were used. A MaxiClean 1400 ultrasonic bath (Unique, São Paulo, Brazil) was employed.

### Standard solutions

SIM RS stock solution at 200 µg mL$^{-1}$ was prepared by accurately weighing 20 mg of SIM in 100 mL volumetric flask, followed by dilution in methanol (50 mL), sonication for 10 min and addition of the same solvent to complete the volume. Aliquots of 0.5, 2.0, 5.0, 7.5 and 10.0 mL were transferred from a precise burette to volumetric flasks (25 mL) in order to obtain SIM standard solutions at 4, 20, 40, 60 and 80 µg mL$^{-1}$ (n = 3) in methanol. All solutions were filtered before the injections (n = 3) in the chromatograph.

### Liquid chromatography

Separation was performed in a RP-8 non-endcapped (LiChroCART® 250-4 LiChrospher® 100, 5 µm, Merck Darmstadt, Germany) column maintained at 30 °C, using methanol:0.1% phosphoric acid (80:20 v/v) as mobile phase with a flow rate of 1.5 mL min$^{-1}$. The backpressure was kept about 125 bar. UV detector was set at $\lambda$ 238 nm. The injection volumes were 10 µL, unless mentioned otherwise.[39-41]

### Analytical curves

SIM standard solutions (triplicate) either in variable injection volumes of 1, 5, 10, 15 and 20 µL from a 40 µg mL$^{-1}$ solution in methanol or in five concentrations were used to build the analytical curves (two or one day) in the ranges of 40-800 ng and 4-80 µg mL$^{-1}$ (independent variable, X) *vs.* SIM chromatographic peak areas (dependent variable, Y, expressed in mAU).[7,15]

### Statistical analysis

Data analyses were performed using the statistical R environment according to its functions.[42] Increasingly complex models were proposed from the simplest linear (*lm*) straight line using OLSM (*lm(y~x)*), and then, using WLSM (*lm(y~x,weights)*).[18] The quadratic model is given as follows.

$$Y_i = \beta_0 + \beta_L X_i + \beta_Q X_i^2 + \varepsilon_i,\ \ i = 1,\ 2, ..., n_T \qquad (1)$$

where $n_T$ is the total number of observations used to estimate the regression coefficients $\beta_0$, $\beta_L$ and $\beta_Q$ (for respective models), and ε is the model error.

The outliers were evaluated using the standardized (also called studentized) residuals (*rstudent( )*). Observations whose residues were greater than 3.0 were considered outliers and then removed (limited to 22% of the dataset) from the dataset.[18,20,43]

The Shapiro-Wilk (*shapiro.test( )*) and Levene (*leveneTest( )*) tests were used to check the assumptions of the model error related to normality and homoscedasticity, respectively.[16,18,19] In addition, the normal quantile-quantile plot (Q-Q plot, *qq.norm( )*) and Bartlett test (*bartlett.test( )*) were also used to verify normality and homoscedasticity, respectively.[19,20] The goodness-of-fit of the model was evaluated by the coefficient of determination (R$^2$) and the mean quadratic error of prediction (MEP),[20] and by observing the residuals *vs.* mass or concentration plot.[13,18] The predictions for the masses (or concentrations), which are obtained from the chromatographic responses through the model equation, were evaluated using the mean relative error (MRE) defined in equation 2 by

$$\text{MRE} = \frac{\sum_{i=1}^{n} \left| x_i - \hat{x}_{-i} \right| / x_i}{\text{n}}, \qquad (2)$$

where $x_i$ is the i-th observation for the mass (or concentration) and $\hat{x}_{-i}$ is the prediction for $x_i$ using the model adjusted to all data, except for the i-th observation. Similar to MEP, the idea of using MRE is to have a global measure for the prediction error (in relative terms) of the model. Measurements similar to MRE are also used to evaluate the goodness-of-fit of regression models.[44]

The R script to adjust and check the quadratic model in equation 1 for a single day, using WLSM, is presented in Figure S1 in the Supplementary Information (SI) section.

Parallelism for SIM mass obtained in the two-day dataset was verified by the addition of two terms to equation 1 resulting in the following:

$$\begin{aligned} Y_i = \beta_0 + \beta_L X_i + \beta_Q X_i^2 + \beta_D D_i + \beta_{DL}(D_i \times X_i) + \\ \beta_{DQ}(D_i \times X_i^2) + \varepsilon_i, \qquad i = 1, 2, ..., n_T \end{aligned} \qquad (3)$$

where D is an indicator for the day of observation (if observation is done on day 1, D = 0; if done on day 2, D = 1). Taking $\beta_D = 0$ and $\beta_{DL} = 0$ in the linear model means that the curves are the same for both days. If only $\beta_{DL} = 0$, the curves are parallel but they have different intercepts. For the quadratic model, the curves are the same if $\beta_D = \beta_{DL} = \beta_{DQ} = 0$ and are parallel if $\beta_{DL} = \beta_{DQ} = 0$. The R script used to adjust and check the linear version of equation 3 with WLSM is depicted in Figure S2 (in the SI section). The results were statistically significant if correspondent *p*-values were less than 0.05. The weights were calculated according to the algorithm shown in Figure S3 (in the SI section).

## Results

The data of the analytical curves obtained by injecting variable volumes (in days 1 and 2) or a fixed volume with different concentrations (in one day) in terms of the values of mean, relative standard deviation (RSD in %) and variance (σ$^2$) are shown in Table 1. The residual plots for SIM masses (ng) and SIM concentrations (µg mL$^{-1}$) after adjusting the ordinary linear model to the dataset are depicted in Figure 1.

The regression models to determine SIM (mass or concentration) calculated by OLSM or WLSM (after outliers removal, if necessary) that showed statistically significant results are presented in Table 2. Q-Q plots for the residuals of the non-weighted and weighted models are shown in Figure 2.

A scatter plot of the variance and mean response calculated at each level of SIM mass (or concentrations) using the data in Table 1 is exhibited in Figure 3. The residual plots for SIM masses (ng) after adjusting the weighted linear model to the dataset obtained for both days and for SIM concentrations (µg mL$^{-1}$) after adjusting the weighted quadratic model are depicted in Figure 4.

## Discussion

The identification and removal of the outliers are important because they inflate the estimated variance,

**Table 1.** Results of peak area (Y) as a function of the variable (mass) or fixed (concentration) volumes of SIM injected in the chromatographic system[a]

| | X | Y, peak area / mAU | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Day 1 | | | | Day 2 | | | |
| | | Value[b] | Mean | RSD / % | $\sigma^2$ | Value[b] | Mean | RSD / % | $\sigma^2$ |
| 40 µg mL$^{-1}$ SIM / µL | 1 | 92.1; 90.7; 91.6 | 91.4 | 0.77 | 0.5 | 93.8; 92.2; 92.4 | 92.8 | 1.29 | 0.7 |
| | 5 | 455.3; 451.4; 453.4 | 453.4 | 0.44 | 3.8 | 457.9; 453.6; 456.0 | 455.9 | 0.51 | 4.6 |
| | 10 | 906.1; 898.4; 901.9 | 902.1 | 0.43 | 14.8 | 909.2; 902.1; 906.4 | 905.9 | 0.38 | 12.7 |
| | 15 | 1359.9; 1342.4; 1347.9 | 1350.1 | 0.66 | 80.0 | 1361.1; 1351.4; 1358.1 | 1357.4 | 0.42 | 24.6 |
| | 20 | 1814.5; 1786.6; 1794.3 | 1798.5 | 0.80 | 207.6 | 1811.2; 1800.9; 1809.0 | 1806.0 | 0.25 | 29.4 |
| SIM[c] / (µg mL$^{-1}$) | 4 | 89.9; 89.3; 90.5 | 89.9 | 0.69 | 0.3 | | | | |
| | 20 | 451.9; 451.6; 448.0 | 450.5 | 0.48 | 4.7 | | | | |
| | 40 | 909.8; 906.4; 911,4 | 909.2 | 0.28 | 6.5 | | | | |
| | 60 | 1364.8; 1380.2; 1385.3 | 1376.8 | 0.78 | 113.9 | | | | |
| | 80 | 1842.1; 1826.3; 1815.4 | 1827.9 | 0.73 | 180.2 | | | | |

[a]Chromatographic conditions: RP-8 (250 × 4 mm, 5 µm) column, 30 °C, λ 238 nm, methanol:0.1% phosphoric acid (80:20 v/v), 1.5 mL min$^{-1}$; [b]each value represents the average of three injections; [c]injection of 10 µL; RSD: relative standard deviation; $\sigma^2$: variance.

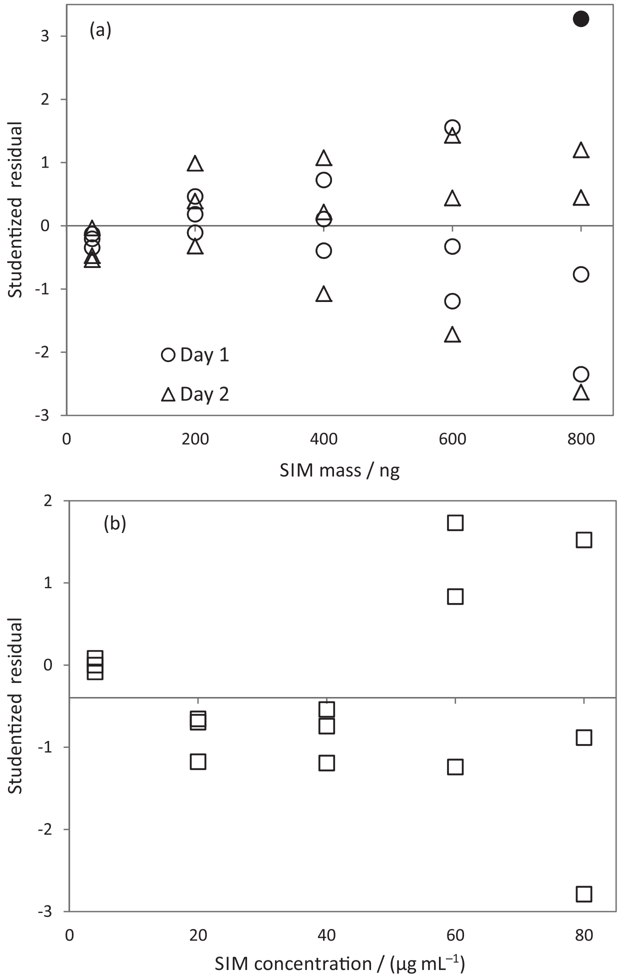


**Figure 1.** Studentized residual plots for SIM (a) mass (40-800 ng) and (b) concentrations (4-80 µg mL$^{-1}$) after adjusting the linear model using OLSM. Chromatographic conditions as in Table 1. Full black circle is an outlier.

increase the probability of type II error (accept the null hypothesis as true when it is false), influence the significance tests for the model parameters and frequently cause violation of the assumption of constant variance error.[18] The simple linear model was initially proposed for the SIM dataset assayed by HPLC-UV (Table 1). Studentized residual plots allowed the identification of one outlier (point 5 in bold, Table 1), which represented 6.6% of the dataset obtained by the injection of variable volumes of 40 µg mL$^{-1}$ SIM on day 1.

The studentized residual plots in Figure 1 tended to exhibit a conical behavior for either SIM mass or concentration, as described for SIM bioanalytical procedures.[22] This behavior was clearly evidenced by the increase in residual variance as the level of SIM mass (or concentration) increased, as can be seen in Table 1. Since the response of SIM mass (or concentration) was positively correlated, the residual variance increased when the response mean increased, which violated the assumption of homoscedasticity.

After the removal of outliers (if necessary), the dataset adjusted to linear models yielded adequate values of R$^2$ (> 0.999) and RSD (< 1.0%), greater than 0.99986 and less than 0.85%, respectively.[10,11] The calculated *p*-values (> 0.06 and > 0.18, respectively) obtained by the Shapiro-Wilk and Levene tests did not allow identification of the problems of normality and homoscedasticity when the linear model was applied to the three datasets. However, problems were detected by the Bartlett test for the homogeneity of variances in the dataset regarding SIM mass for day 1 and SIM concentrations (Table 2). The Q-Q plot of the linear model showed points

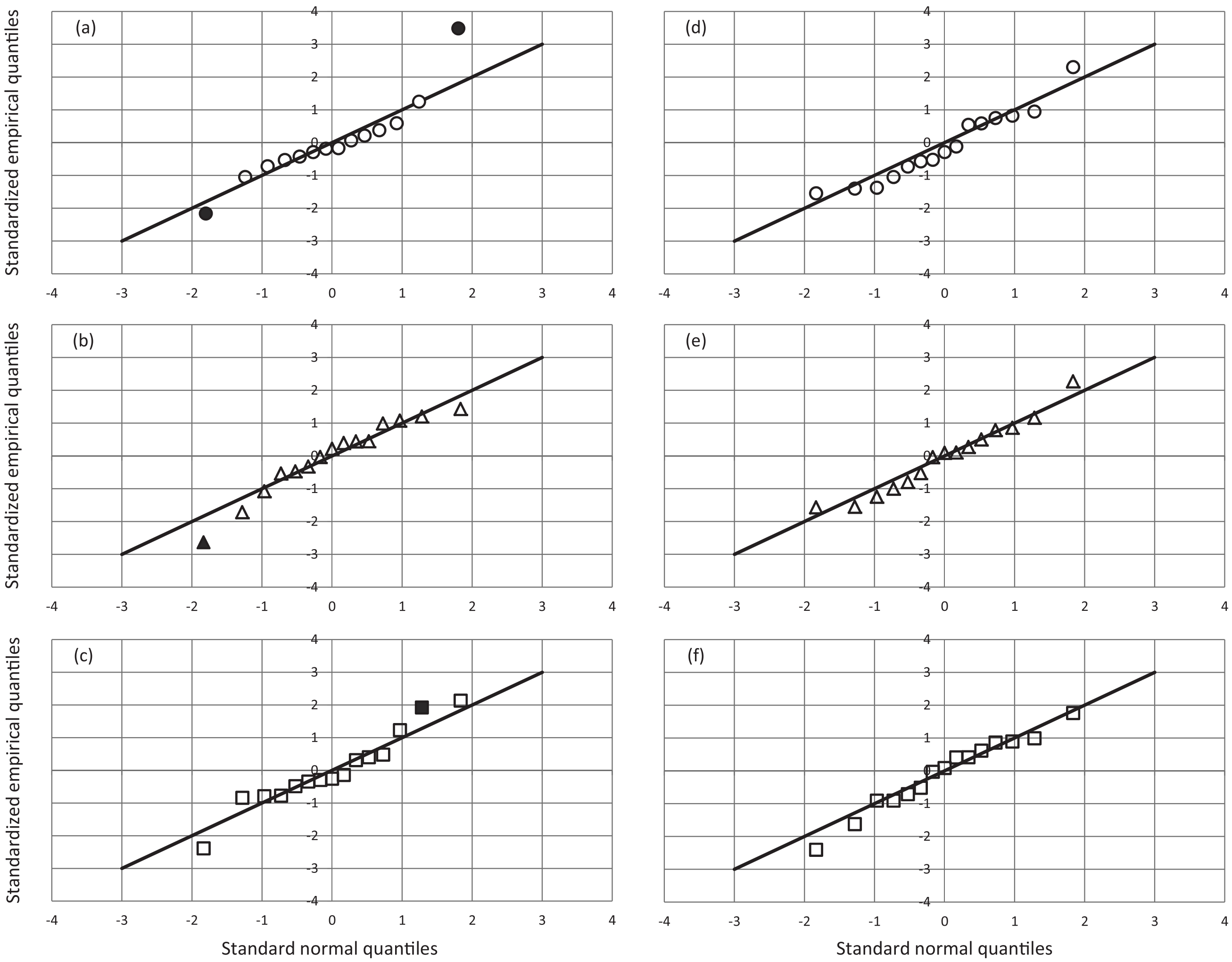


**Figure 2.** SIM residual Q-Q plots by HPLC-UV injection of variable volumes (n = 5) on day 1 and on day 2, and for fixed volume (10 µL) of five concentrations on day 1, obtained from OLSM-adjusted linear models (a, b and c); and obtained from WLSM-linear (d and e) and WLSM-quadratic (f) models. Points outside of the confidence limits are in black. Chromatographic conditions as in Table 1.

**Table 2.** Results obtained for linear and quadratic regression models calculated by OLSM or WLSM for SIM determination using HPLC[a]

| SIM range | Model | Weight[b] | Equation | $R^2$ | RSD / % | MEP[c] | MRE[d] | $N_{out}$[e] | Test statistic ($p$-value[f]) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Levene | Bartlett | Shapiro-Wilk |
| 40-800 ng | | | | | | | | | | | |
| Day 1 | linear | 1 | 4.641 + 2.238x | 0.99993 | 0.61 | 61.4 | 0.00819 | 1 | 1.33 (0.33) | 10.3 ($3.6 \times 10^{-2}$) | 0.88 (0.06) |
| | linear | 1/W | 1.555 + 2.253x | 0.99996 | 0.11 | 55.3 | 0.00586 | 0 | 0.14 (0.97) | 0.51 (0.97) | 0.94 (0.40) |
| Day 2 | linear | 1 | 3.715 + 2.255x | 0.99997 | 0.37 | 14.0 | 0.00574 | 0 | 0.69 (0.62) | 5.82 (0.21) | 0.94 (0.35) |
| | linear | 1/W | 2.939 + 2.258x | 0.99998 | 0.11 | 13.7 | 0.00504 | 0 | 0.23 (0.92) | 0.93 (0.92) | 0.96 (0.74) |
| 4-80 µg mL$^{-1}$ | | | | | | | | | | | |
| | linear | 1 | –4.760 + 22.932x | 0.99986 | 0.85 | 75.5 | 0.01327 | 0 | 1.97 (0.18) | 15.2 ($4.3 \times 10^{-3}$) | 0.94 (0.42) |
| | linear | 1/W | –1.601 + 22.742x | 0.99991 | 0.18 | 112.8 | 0.00729 | 0 | 0.29 (0.88) | 2.85 (0.58) | 0.92 (0.17) |
| | quadratic | 1/W | 0.206 + 22.374x + 0.008x$^2$ | 0.99996 | 0.12 | 119.6 | 0.00621 | 0 | 0.23 (0.92) | 2.06 (0.72) | 0.97 (0.85) |

[a] Chromatographic conditions as in Table 1; [b] W: weights calculated as in Figure S3 (in the SI section); [c] MEP: mean quadratic error of prediction; [d] MRE: mean relative error; [e] $N_{out}$: number of outliers; [f] statistically significant if less than 0.05.

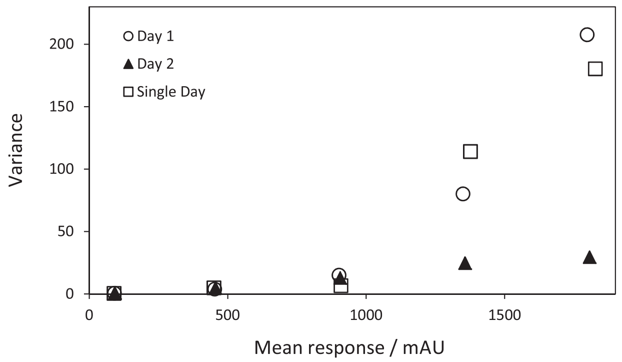


**Figure 3.** Estimates of the variance response ($\hat{\sigma}^2_Y$) *vs.* estimates of mean response ($\hat{\mu}_Y$) at each level of SIM masses (for days 1 and 2) and SIM concentrations (single day). Chromatographic conditions as in Table 1.

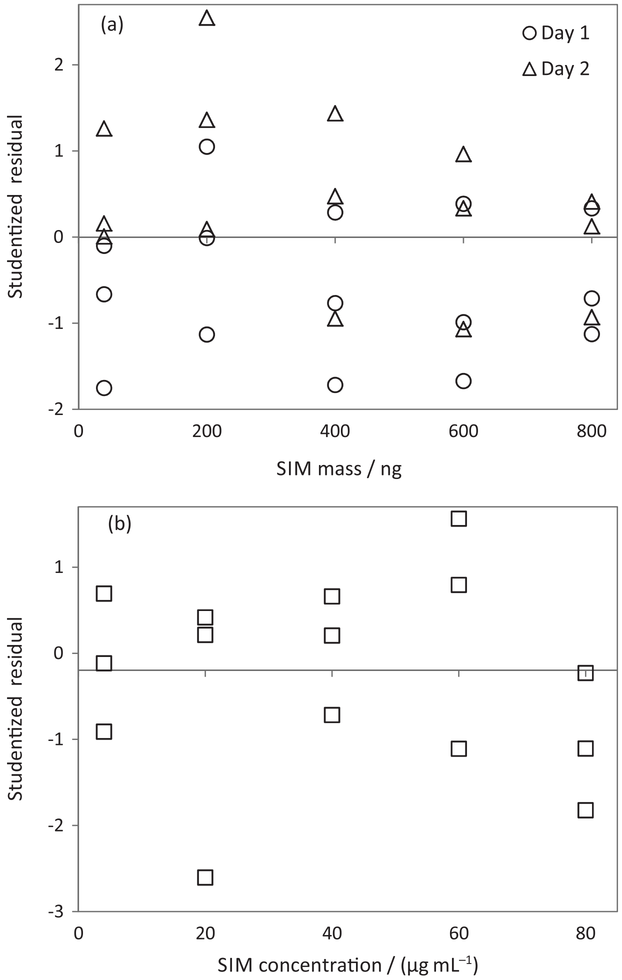


**Figure 4.** Studentized residual plots obtained for SIM (a) mass (40-800 ng) for two days and (b) concentrations (4-80 µg mL$^{-1}$), after adjusting the weighted linear and quadratic models, respectively. Chromatographic conditions as in Table 1.

outside (in black) the confidence interval for all datasets, as in Figures 2a, 2b and 2c. Thus, the linear models could not meet the OLSM assumptions, which required other models to fit the dataset.

The quadratic model calculated by OLSM was not statistically significant ($p$-value ≥ 0.53) when applied to the dataset obtained either from SIM mass for day 2 (after removal of one outlier) or SIM concentrations. Although the values of R$^2$ (> 0.99995) and RSD (< 0.54%) were appropriate, greater than 0.999 and less than 1.0%, respectively, the quadratic model calculated by OLSM was statistically inconclusive ($p$-value ≥ 0.057) when applied to the dataset obtained from SIM mass for day 1 (after removal of one outlier).[10,11] Furthermore, the quality of the method was inadequate due to the lack of normality and the homoscedasticity of the model errors, as verified by the $p$-values of the Shapiro-Wilk (0.01) and Barlett (0.03) tests, respectively. The normality problem was confirmed by a corresponding Q-Q plot, which was similar to that mentioned in Figure 2a.

Independent on the applied model (linear or quadratic) and the dataset, there was at least one problem related to the quality of the method (homoscedasticity and/or normality). These variance problems are well known requirements for not using OLSM.[18,19]

A possible solution to these problems is to use weights in the estimation of the model coefficients, when WLSM is advised for heteroscedasticity cases.[18] The most common weighting factors reported for SIM in bioanalytical procedures are 1/x, 1/x$^2$ and 1/y.[23-26] Alternatively, the variance of the responses can be leveled by transforming the response, for instance, by using the Box-Cox method.[21]

Since calculation of weights is a very important step for the success of the model adjustment, the relationship between the variance and mean of the responses for each SIM mass and concentration level was examined using scatter plots for all datasets. The aim was to predict an appropriate weighting factor to make residual variance more homogeneous.

In this case, the weights were defined as the inverse of the variance of the observations at the determined level of SIM mass (or concentration). Thus, observations taken at levels with larger variances have smaller weights.

Since the variance of the responses is related to their mean, one can adjust a linear model to estimate the variance of the responses through the mean of the responses. The relationship between the variances and means was clearly nonlinear (Figure 3). Residual variance seemed to be an exponential function of the mean of the responses for all dataset to a greater (day 1, SIM mass and SIM concentration) or lesser (day 2, SIM mass) degree. Therefore, the relationship between the natural logarithm of the variance and the mean would be linear, and as such, a linear regression model was adjusted to estimate the variances and weights. The dependent and independent variables were the natural logarithm of the variance of the responses and the mean of the responses, respectively. Hence, a weight was estimated for each level of SIM mass (or concentration). These weights were assigned to the three observations of each level of SIM mass (or concentration). Therefore, in order to correct the problems of the homogeneity of variance, the reciprocal of the variance of responses was used as a weighting factor for the linear and quadratic models. It was calculated as depicted in R script (Figure S3 in the SI section).

Initially, a weighted linear model was adjusted to the three datasets and no outlier was observed by the treatment of the studentized residuals (Table 2). The removal of outliers is not an ideal condition, since it reduces the scarce degrees of freedom of the sum of squared errors even more. The residual plots adjusted by the weighted linear models did not exhibit any conical behavior; otherwise, a random distribution around the value zero without any trend for SIM concentration (data not shown) and SIM mass datasets for days 1 and 2 is exhibited (Figure 4a). All models showed more adequate values of $R^2$ (> 0.99991) and RSD (< 0.18) than those obtained with the respective non-weighted linear models (Table 2), thereby indicating the quality of the model.[10,11,18] For all datasets, the *p*-values of the Shapiro-Wilk and Levene tests were greater than 0.17 and 0.88, respectively, considering the datasets obtained from SIM mass, which were also more appropriate than those observed when the linear model was adjusted using OLSM. The *p*-values of the Bartlett test confirmed the adequate homogeneity of the variance for all datasets. Q-Q plots showed adequate distribution of the residues. The results were better than those observed when the simplest linear model was adjusted to the datasets obtained from SIM mass, as in Figures 2d and 2e.

Last, the fitting of the weighted quadratic model was only statistically significant ($2.48 \times 10^{-13}$) for the dataset obtained from SIM concentration. The residual plot showed minor discrepancy in the variances, suggesting homogeneity, as in Figure 4b. This result was confirmed by the adequate *p*-values obtained from the Levene (0.92) and Bartlett (0.72) tests. The best values of $R^2$ and RSD were observed using the weighted quadratic model (Table 2). The method also exhibited appropriate normality evaluated by the Shapiro-Wilk test and the Q-Q plot, as shown in Figure 2f.

It is worth noting that Q-Q plots for the residuals of non-weighted models revealed problems in the assumption of normality, while the Q-Q plots for the weighted models (right column) did not show any such problems.

Further, the models using WLSM showed smaller MRE values than those obtained using OLSM (Table 2). This means that the errors in the predictions of mass (or concentration) were smaller when using WLSM than when using OLSM to estimate the regression equation.

Comparing the weighted linear equations applied to SIM mass determination on days 1 and 2 by ANOVA, no statistical difference was found between the beta coefficients, which means that the parallelism was attested (Figure S2). Thus, a single weighted linear curve expressed by $y = 1.885 + 2.285x$ with $R^2$ 0.99996 and RSD 0.12% could be used to determine SIM mass. This dataset also showed adequate normality (Shapiro-Wilk *p*-value 0.59) and homoscedasticity (Levene and Bartlett *p*-values 0.36 and 0.81, respectively).

In this study, the use of mass or concentration in the construction of analytical curves seemed to influence the model, as long as linear and quadratic fittings were, respectively, observed. Independent on the mass or concentration, it was observed that WLSM was adequate for curve fitting, unlike OLSM, due to the wide employed interval (which varied between 10 and 200% of the analytical range) as it had been intentionally designed for future drug dissolution studies.[13,14] Furthermore, the suitability of WLSM was previously suggested by the removal of less outlier observations in comparison to OLSM. The difference in the models can be explained by the error associated with the standard dilutions. This error was more relevant when working with distinct concentrations, as in the present case expressed by quadratic fitting. Under variable volumes, the dilution error was less significant as it was substituted by the injection error, exemplified in the case of linear fitting.

## Conclusions

As long as the good laboratory practice and validation criteria are fulfilled, the most adequate model to express the relationship between mass or concentration and its response should be selected considering the assumptions of the least squares method related to the variance of the residues. The possibility of the mass or concentration influencing models differently needs to be noted. Generally, the simplest linear model adjusted by OLSM should preferably be adopted. However, when OLSM does not show adequacy to express the relationship between variables, the source of heteroscedasticity should be investigated before leveling the response variances by WLSM. Then, a question arises as to what can be an alternative for the adjustment of the model. In this case, the choice of an appropriate weighting factor is a better alternative than using response transformations (inversion, square root, square inversion, etc.) since they can make the model more difficult to both interpret and apply.

Therefore, the analyst should be careful when making a choice and avoid misleading interpretations of data.

## Supplementary Information

Supplementary data (R scripts) are available free of charge at http://jbcs.sqt.org.br as a PDF file.

## Acknowledgments

Fapemig and Capes for funding sources.

## References

1. Thompson, M.; Ellison, S. L. R.; Wood, R.; *Pure Appl. Chem.* **2002**, *74*, 835.
2. International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use; *Validation of Analytical Procedures: Text and Methodology Q2 (R1)*, Geneva, 2005.
3. US Department of Health and Human Services; US Food and Drug Administration (US FDA); *Guidance for Industry: Analytical, Procedures and Methods Validation (Draft guidance)*, Rockville, USA, 2000.
4. EURACHEM Group; *Fitness for Purpose of Analytical Methods: a Laboratory Guide to Method Validation and Related Topics*; LGC: Teddington, UK, 1998, http://www.eurachem.org/images/stories/Guides/pdf/valid.pdf, accessed in March 2013.
5. American Organization of Analytical Chemists (AOAC); *Peer Verified Methods Program, Manual on Policies and Procedures*, AOAC International: Arlington, USA, 1998.
6. United States Pharmacopeia Convention, *The United States Pharmacopoeia*, 34th ed.: Rockville, USA, 2011, ch. 1225.
7. British Pharmacopoeia, Her Majesty's Stationery Office: London, 2011, ch. IIIE.
8. Brazilian National Health Surveillance Agency (ANVISA); *Guide for the Validation of Analytical and Bioanalytical Methods*, Resolution No. 899, May 29, 2003, Diário Oficial: Brasília, Brazil, 2003.
9. Brazilian National Institute of Metrology, Quality and Technology (INMETRO); *Guidance on Validation of Methods of Chemical Tests*, DOQ-CGCRE-008, Brasília, Brazil, 2007.
10. Ermer, J.; *J. Pharm. Biomed. Anal.* **2001**, *24*, 755.
11. Shabir, G. A.; *J. Chromatogr., A* **2003**, *987*, 57.
12. Rozet, E.; Ceccato, A.; Hubert, C.; Ziemons, E.; Oprean, R; Rudaz, S.; Boulanger, B.; Hubert, P.; *J. Chromatogr., A* **2007**, *1158*, 111.
13. González, A. G.; Herrador, M. A.; *TrAC: Trends Anal. Chem.* **2007**, *26*, 227.
14. Hubert, Ph.; Nguyen-Huu, J. J.; Boulanger, B.; Chapuzet, E.; Chiap, P.; Cohen, N.; Compagnon, P.-A.; Dewé, W.; Feinberg, M.; Lallier, M.; Laurentie, M.; Mercier, N.; Muzard, G.; Nivet, C.; Valat, L.; Rozet, E.; *J. Pharm. Biomed. Anal.* **2007**, *45*, 70.
15. Araujo, P.; *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2009**, *877*, 2224.
16. Hubert, Ph.; Nguyen-Huu, J.-J.; Boulanger, B.; Chapuzet, E.; Cohen, N.; Compagnon, P.-A.; Dewé, W.; Feinberg, M.; Laurentie, M.; Mercier, N.; Muzard, G.; Valat, L.; Rozet, E.; *J. Pharm. Biomed. Anal.* **2007**, *45*, 82.
17. Rozet, E.; Marini, R. D.; Ziemons, E.; Boulanger, B.; Hubert, P.; *J. Pharm. Biomed. Anal.* **2011**, *55*, 848.
18. Meloun, M.; Militký, J.; Kupka, K.; Brereton, R. G.; *Talanta* **2002**, *57*, 721.
19. de Souza, S. V. C.; Junqueira, R. G.; *Anal. Chim. Acta* **2005**, *552*, 25.
20. Meloun, M.; Militký, J.; *Anal. Chim. Acta* **2001**, *439*, 169.
21. Singtoroj, T.; Tarning, J.; Annerberg, A.; Ashton, M.; Bergqvist, Y.; White, N. J.; Lindegardh, N.; Day, N. P. J.; *J. Pharm. Biomed. Anal.* **2006**, *41*, 219.
22. Almeida, A. M.; Castel-Branco, M. M.; Falcão, A. C.; *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2002**, *774*, 215.
23. Barrett, B.; Huclová, J.; Bořek-Dohalský, V.; Němeca, B.; Jelínek, I.; *J. Pharm. Biomed. Anal.* **2006**, *41*, 517.
24. Selvan, P. S.; Pal, T. K.; *J. Pharm. Biomed. Anal.* **2009**, *49*, 780.
25. Apostolou, C.; Kousoulos, C.; Dotsikas, Y.; Soumelas, G.-S.; Kolocouri, F.; Ziaka, A.; Loukas, Y. L.; *J. Pharm. Biomed. Anal.* **2008**, *46*, 771.
26. Yang, H.; Feng, Y.; Luan, Y.; *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2003**, *785*, 369.
27. Gomes, F. P.; García, P. L.; Alves, J. M. P.; Singh, A. K.; Kedor-Hackmann, E. R. M.; Santoro, M. I. R. M.; *Lat. Am. J. Pharm.* **2009**, *28*, 261**.**
28. Abu-Nameh, E. S. M.; Shawabkeh, R. A.; Ali, A.; *J. Anal. Chem.* **2006**, *61*, 63.
29. Godoy, R.; Godoy, C. G.; Diego, M.; Gómez, C.; *J. Chil. Chem. Soc.* **2004**, *49*, 289.
30. Markman, B. E. O.; Rosa, P. C. P.; Koschtschak, M. R. W.; *Rev. Saúde Publica* **2010**, *44*, 1055.
31. Ashfaq, M.; Khan, I. U.; Asghar, M. N.; *J. Chil. Chem. Soc.* **2008**, *53*, 1617.
32. Ali, H.; Nazzal, S.; *J. Pharm. Biomed. Anal.* **2009**, *49*, 950.
33. Hefnawy, M.; Al-Omar, M.; Julkhuf, S.; *J. Pharm. Biomed. Anal.* **2009**, *50*, 527.
34. Kumar, V.; Shah, R. P.; Singh, S.; *J. Pharm. Biomed. Anal.* **2008**, *47*, 508.
35. Hsu, K.-H.; Chen, C.; *Measurement* **2010**, *43*, 1525.
36. César, I. C.; Braga, F. C.; Vianna-Soares, C. D.; Nunan, E. A.; Pianetti, G. A.; Condessa, F. A.; Barbosa, T. A. F.; Campos, L. M. M.; *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2006**, *836*, 74.

37. Melo, A. C.; Cota, B. B.; Oliveira, A. B.; Braga, F. C.; *Fitoterapia* **2001**, *72*, 40.
38. Wu, T.; Bligh, S. W. A.; Gu, L.-h.; Wang, Z.-t.; Liu, H.-p.; Cheng, X.-m.; Branford-White, C. J.; Hu, Z.-b.; *Fitoterapia* **2005**, *76*, 157.
39. Marques-Marinho, F. D.; Zanon, J. C. C.; Sakurai, E.; Reis, I. A.; Lima, A. A.; Vianna-Soares, C. D.; *Braz. J. Pharm. Sci.* **2011**, *47*, 495.
40. Marques-Marinho, F. D.; dos Santos, A. L.; Zanon, J. C. C.; Reis, I. A.; Lima, A. A.; Vianna-Soares, C. D.; *Quim. Nova* **2012**, *35*, 1233.
41. Marques-Marinho, F. D.; Freitas, B. D.; Zanon, J. C. C.; Reis, I. A.; Lima, A. A.; Vianna-Soares, C. D.; *Curr. Pharm. Anal.* **2013**, *9*, 2.
42. The R Foundation for Statistical Computing, Austria, 2010, ISBN 3-900051-07-0, http://www.R-project.org, accessed in March 2013.
43. Horwitz,W.; *Pure Appl. Chem.* **1995**, *67*, 331.
44. http://www.dsr.inpe.br/sbsr2011/files/p0342.pdf, accessed in August 2013.

*Submitted: June 16, 2013*
*Published online: August 13, 2013*

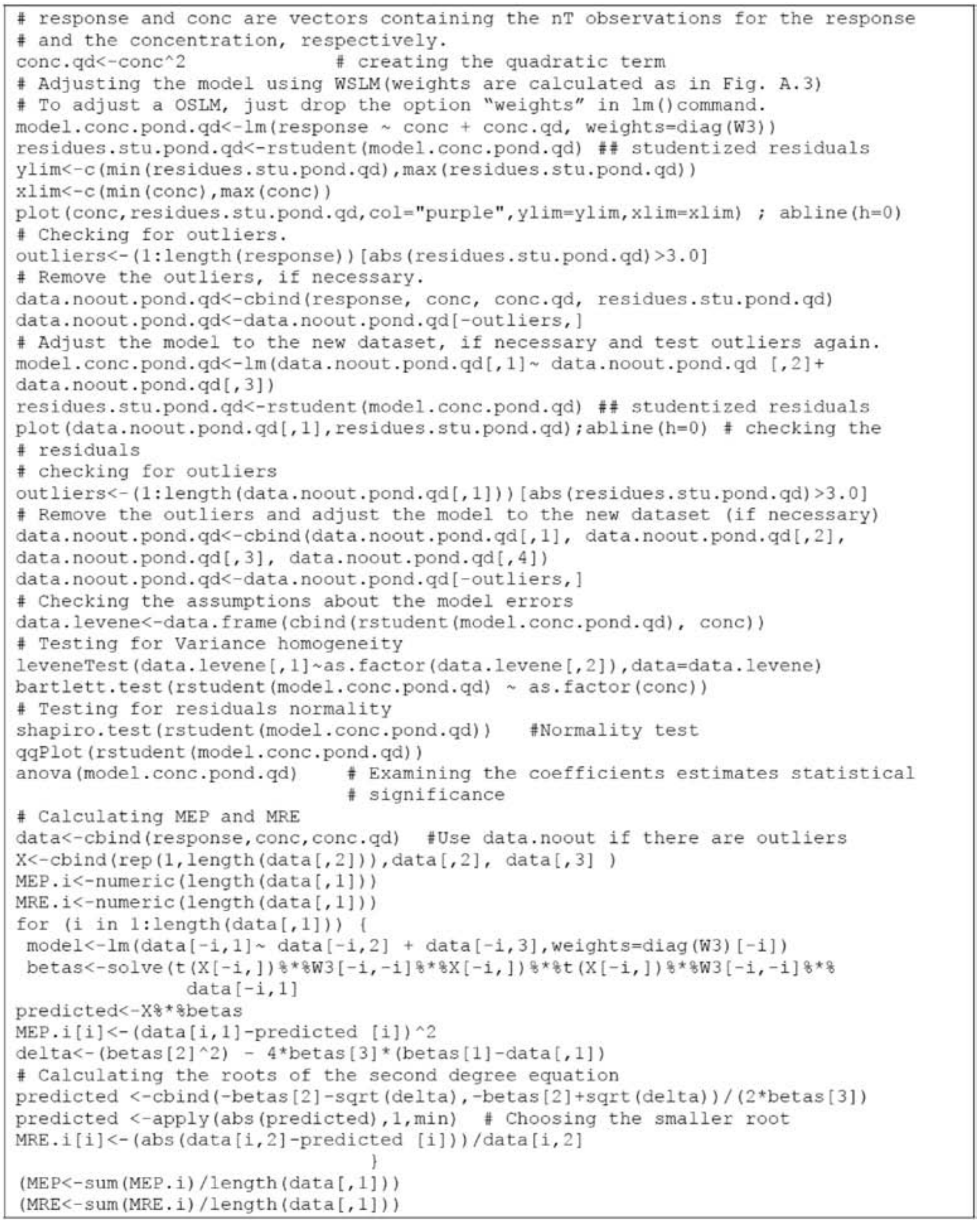
*J. Braz. Chem. Soc.*, Vol. 24, No. 9, S1-S3, 2013.
Printed in Brazil - ©2013 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

*Supplementary Information*

# Construction of Analytical Curve Fit Models for Simvastatin using Ordinary and Weighted Least Squares Methods

*Flávia D. Marques-Marinho,*[*,a] Ilka A. Reis[b] and Cristina D. Vianna-Soares[a]*

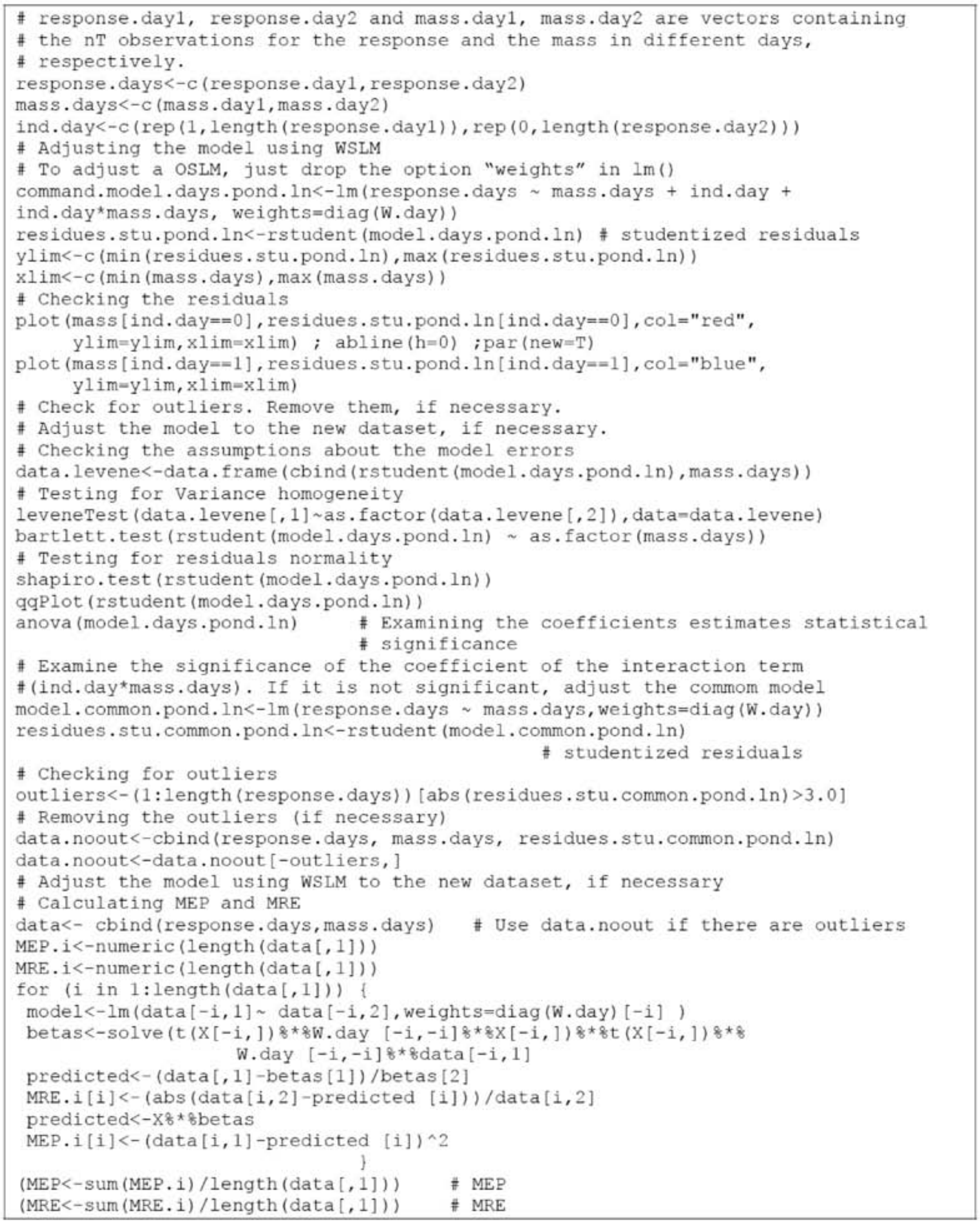
*[a]Department of Pharmaceutical Products, Faculty of Pharmacy and [b]Department of Statistics,
Institute of Exact Sciences, Federal University of Minas Gerais, Av. Pres. Antônio Carlos, 6627,
31270-901 Belo Horizonte-MG, Brazil*



```
# response and conc are vectors containing the nT observations for the response
# and the concentration, respectively.
conc.qd<-conc^2             # creating the quadratic term
# Adjusting the model using WSLM(weights are calculated as in Fig. A.3)
# To adjust a OSLM, just drop the option “weights” in lm()command.
model.conc.pond.qd<-lm(response ~ conc + conc.qd, weights=diag(W3))
residues.stu.pond.qd<-rstudent(model.conc.pond.qd) ## studentized residuals
ylim<-c(min(residues.stu.pond.qd),max(residues.stu.pond.qd))
xlim<-c(min(conc),max(conc))
plot(conc,residues.stu.pond.qd,col="purple",ylim=ylim,xlim=xlim) ; abline(h=0)
# Checking for outliers.
outliers<-(1:length(response))[abs(residues.stu.pond.qd)>3.0]
# Remove the outliers, if necessary.
data.noout.pond.qd<-cbind(response, conc, conc.qd, residues.stu.pond.qd)
data.noout.pond.qd<-data.noout.pond.qd[-outliers,]
# Adjust the model to the new dataset, if necessary and test outliers again.
model.conc.pond.qd<-lm(data.noout.pond.qd[,1]~ data.noout.pond.qd [,2]+
data.noout.pond.qd[,3])
residues.stu.pond.qd<-rstudent(model.conc.pond.qd) ## studentized residuals
plot(data.noout.pond.qd[,1],residues.stu.pond.qd);abline(h=0) # checking the
# residuals
# checking for outliers
outliers<-(1:length(data.noout.pond.qd[,1]))[abs(residues.stu.pond.qd)>3.0]
# Remove the outliers and adjust the model to the new dataset (if necessary)
data.noout.pond.qd<-cbind(data.noout.pond.qd[,1], data.noout.pond.qd[,2],
data.noout.pond.qd[,3], data.noout.pond.qd[,4])
data.noout.pond.qd<-data.noout.pond.qd[-outliers,]
# Checking the assumptions about the model errors
data.levene<-data.frame(cbind(rstudent(model.conc.pond.qd), conc))
# Testing for Variance homogeneity
leveneTest(data.levene[,1]~as.factor(data.levene[,2]),data=data.levene)
bartlett.test(rstudent(model.conc.pond.qd) ~ as.factor(conc))
# Testing for residuals normality
shapiro.test(rstudent(model.conc.pond.qd))   #Normality test
qqPlot(rstudent(model.conc.pond.qd))
anova(model.conc.pond.qd)    # Examining the coefficients estimates statistical
                             # significance
# Calculating MEP and MRE
data<-cbind(response,conc,conc.qd)  #Use data.noout if there are outliers
X<-cbind(rep(1,length(data[,2])),data[,2], data[,3] )
MEP.i<-numeric(length(data[,1]))
MRE.i<-numeric(length(data[,1]))
for (i in 1:length(data[,1])) {
 model<-lm(data[-i,1]~ data[-i,2] + data[-i,3],weights=diag(W3)[-i])
 betas<-solve(t(X[-i,])%*%W3[-i,-i]%*%X[-i,])%*%t(X[-i,])%*%W3[-i,-i]%*%
               data[-i,1]
predicted<-X%*%betas
MEP.i[i]<-(data[i,1]-predicted [i])^2
delta<-(betas[2]^2) - 4*betas[3]*(betas[1]-data[,1])
# Calculating the roots of the second degree equation
predicted <-cbind(-betas[2]-sqrt(delta),-betas[2]+sqrt(delta))/(2*betas[3])
predicted <-apply(abs(predicted),1,min)  # Choosing the smaller root
MRE.i[i]<-(abs(data[i,2]-predicted [i]))/data[i,2]
                                }
 (MEP<-sum(MEP.i)/length(data[,1]))
 (MRE<-sum(MRE.i)/length(data[,1]))
```

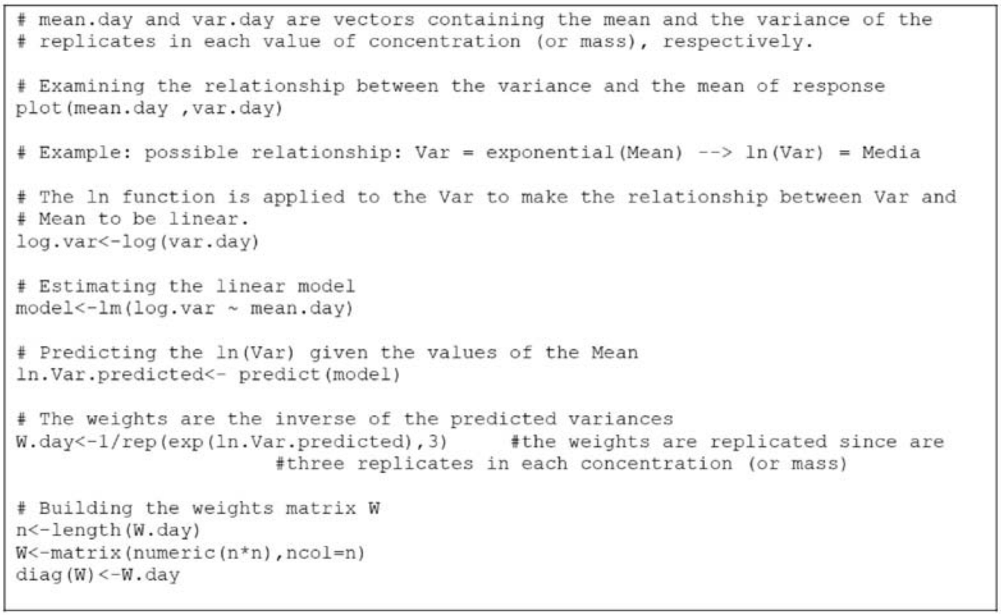
**Figure S1.** R script to adjust and check the quadratic model in equation 1 for a single day using the WLSM when independent variable (X) is expressed as concentration.

*e-mail: flaviadmar@hotmail.com



```
# response.day1, response.day2 and mass.day1, mass.day2 are vectors containing
# the nT observations for the response and the mass in different days,
# respectively.
response.days<-c(response.day1,response.day2)
mass.days<-c(mass.day1,mass.day2)
ind.day<-c(rep(1,length(response.day1)),rep(0,length(response.day2)))
# Adjusting the model using WSLM
# To adjust a OSLM, just drop the option "weights" in lm()
command.model.days.pond.ln<-lm(response.days ~ mass.days + ind.day +
ind.day*mass.days, weights=diag(W.day))
residues.stu.pond.ln<-rstudent(model.days.pond.ln) # studentized residuals
ylim<-c(min(residues.stu.pond.ln),max(residues.stu.pond.ln))
xlim<-c(min(mass.days),max(mass.days))
# Checking the residuals
plot(mass[ind.day==0],residues.stu.pond.ln[ind.day==0],col="red",
     ylim=ylim,xlim=xlim) ; abline(h=0) ;par(new=T)
plot(mass[ind.day==1],residues.stu.pond.ln[ind.day==1],col="blue",
     ylim=ylim,xlim=xlim)
# Check for outliers. Remove them, if necessary.
# Adjust the model to the new dataset, if necessary.
# Checking the assumptions about the model errors
data.levene<-data.frame(cbind(rstudent(model.days.pond.ln),mass.days))
# Testing for Variance homogeneity
leveneTest(data.levene[,1]~as.factor(data.levene[,2]),data=data.levene)
bartlett.test(rstudent(model.days.pond.ln) ~ as.factor(mass.days))
# Testing for residuals normality
shapiro.test(rstudent(model.days.pond.ln))
qqPlot(rstudent(model.days.pond.ln))
anova(model.days.pond.ln)      # Examining the coefficients estimates statistical
                               # significance
# Examine the significance of the coefficient of the interaction term
#(ind.day*mass.days). If it is not significant, adjust the commom model
model.common.pond.ln<-lm(response.days ~ mass.days,weights=diag(W.day))
residues.stu.common.pond.ln<-rstudent(model.common.pond.ln)
                                                # studentized residuals
# Checking for outliers
outliers<-(1:length(response.days))[abs(residues.stu.common.pond.ln)>3.0]
# Removing the outliers (if necessary)
data.noout<-cbind(response.days, mass.days, residues.stu.common.pond.ln)
data.noout<-data.noout[-outliers,]
# Adjust the model using WSLM to the new dataset, if necessary
# Calculating MEP and MRE
data<- cbind(response.days,mass.days)   # Use data.noout if there are outliers
MEP.i<-numeric(length(data[,1]))
MRE.i<-numeric(length(data[,1]))
for (i in 1:length(data[,1])) {
 model<-lm(data[-i,1]~ data[-i,2],weights=diag(W.day)[-i] )
 betas<-solve(t(X[-i,])%*%W.day [-i,-i]%*%X[-i,])%*%t(X[-i,])%*%
                    W.day [-i,-i]%*%data[-i,1]
 predicted<-(data[,1]-betas[1])/betas[2]
 MRE.i[i]<-(abs(data[i,2]-predicted [i]))/data[i,2]
 predicted<-X%*%betas
 MEP.i[i]<-(data[i,1]-predicted [i])^2
                                    }
(MEP<-sum(MEP.i)/length(data[,1]))     # MEP
(MRE<-sum(MRE.i)/length(data[,1]))     # MRE
```

**Figure S2.** R script to adjust and check the linear version of the model in equation 3 using the WLSM when independent variable (X) is expressed as mass.



```
# mean.day and var.day are vectors containing the mean and the variance of the
# replicates in each value of concentration (or mass), respectively.

# Examining the relationship between the variance and the mean of response
plot(mean.day ,var.day)

# Example: possible relationship: Var = exponential(Mean) --> ln(Var) = Media

# The ln function is applied to the Var to make the relationship between Var and
# Mean to be linear.
log.var<-log(var.day)

# Estimating the linear model
model<-lm(log.var ~ mean.day)

# Predicting the ln(Var) given the values of the Mean
ln.Var.predicted<- predict(model)

# The weights are the inverse of the predicted variances
W.day<-1/rep(exp(ln.Var.predicted),3)     #the weights are replicated since are
                      #three replicates in each concentration (or mass)

# Building the weights matrix W
n<-length(W.day)
W<-matrix(numeric(n*n),ncol=n)
diag(W)<-W.day
```

**Figure S3.** R script to examine the relationship between the variance and the mean of response and to calculate the weights to be used in WLSM.