Finding best weighted linear regression model for heteroscedastic data by maximizing likelihood function

Vicențiu Taciuc[1], Augustin-Cătălin Moț[1,2]

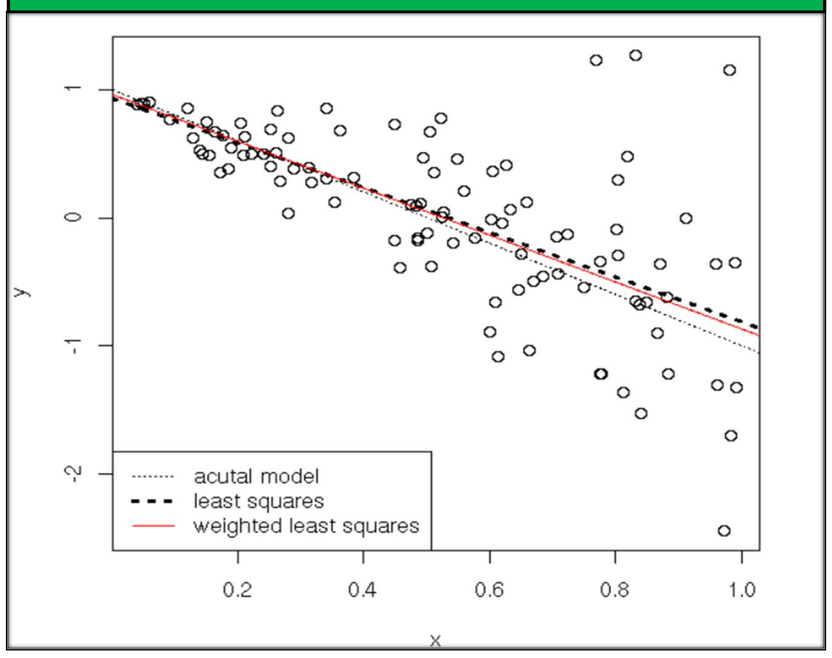
[1]Babes – Bolyai University, Faculty of Chemistry and Chemical Engineering, Department of Chemistry, 11 Arany Janos, 400028, Cluj - Napoca, Romania

[2] National Institute for Research and Development of Isotopic and Molecular Technologies, 67-103 Donath, 400293, Cluj – Napoca, Romania

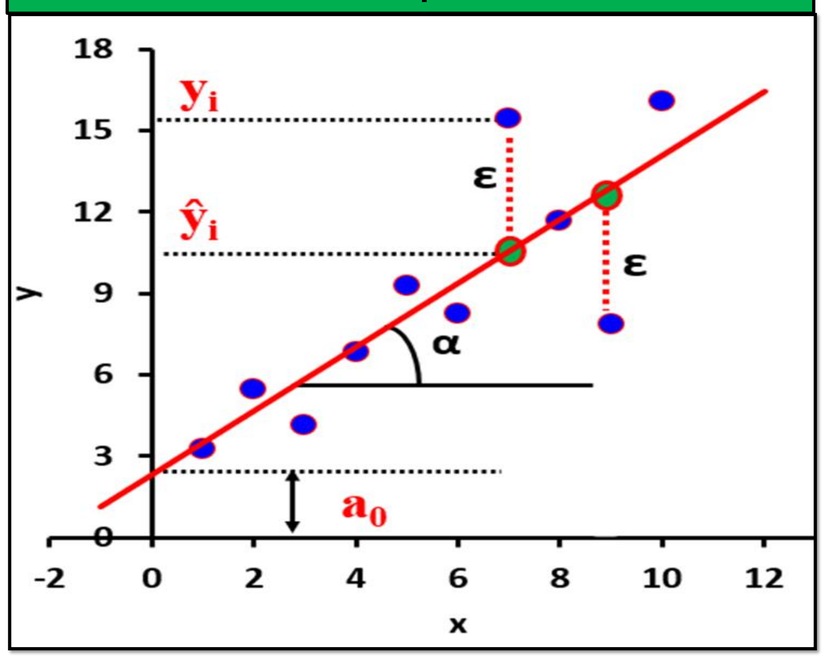
E-mail: taciuc.vicentiu@gmail.com

Abstract: Ordinary least squares regression (OLS) is the most frequently used method applied in analytical chemistry for estimation of the parameters of a calibration curve. Most important assumptions in OLS regression presume a linear data set, little or no error for independent variable, independence and normal distribution of the residuals, no outliers and last but not least constant variance or a homoscedasticity. A dataset whose variance of the residuals depends upon the independent variable is called heteroscedastic as to oppose to homoscedastic ones. Residuals plots or statistical tests such as Breusch-Pagan or White test are commonly used for diagnosing a heteroscedastic behavior. However, these tests are not efficient if the dataset is low in size. If the dataset is proven to be heteroscedastic, weighted linear regression, log transformation or nonparametric median regression could be applied instead. In the context of a weighted linear regression, it is necessary to choose a suitable weight that leads to the best predictive model and most frequently is used a weight like 1/x, 1/x2, 1/x1/2 or generally 1/xγ. In this presentation, a novel way to estimate the best weight for weighted linear regression will be presented using the profile of the log-likelihood regression function. Moreover, this method appears to be a goldfish since not only indicates the most appropriate weight but also diagnose the heteroscedastic profile and variance non-homogeneity along the x axis.

A comparison of different types of linear regression



Least squares method



Parameters of regression

$$a = \frac{\sum y - b\sum x}{n} \qquad b = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

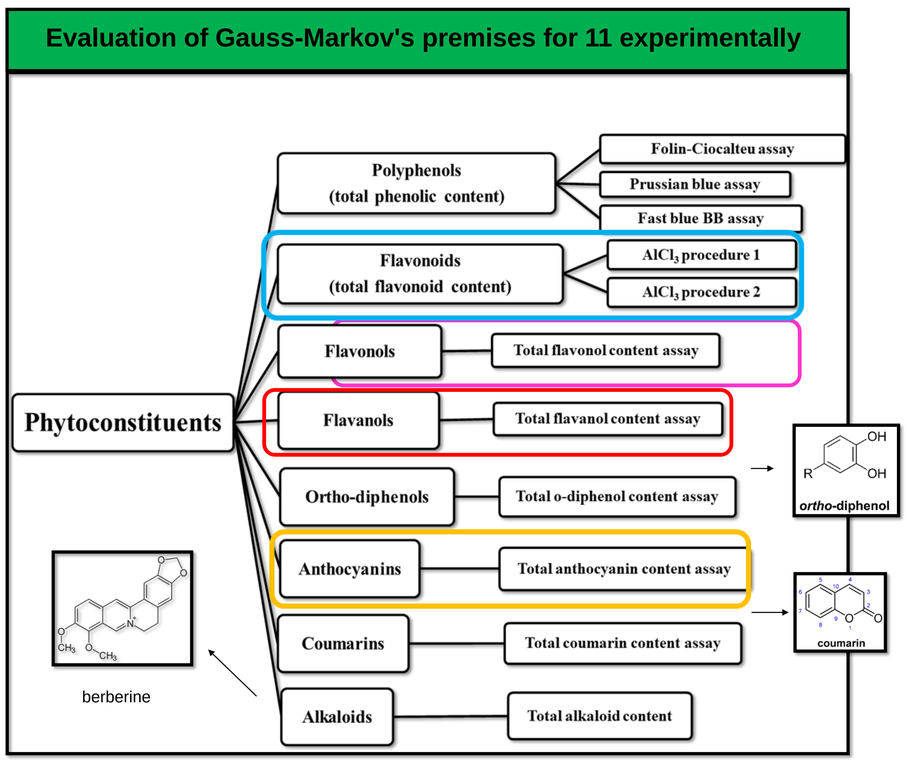
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

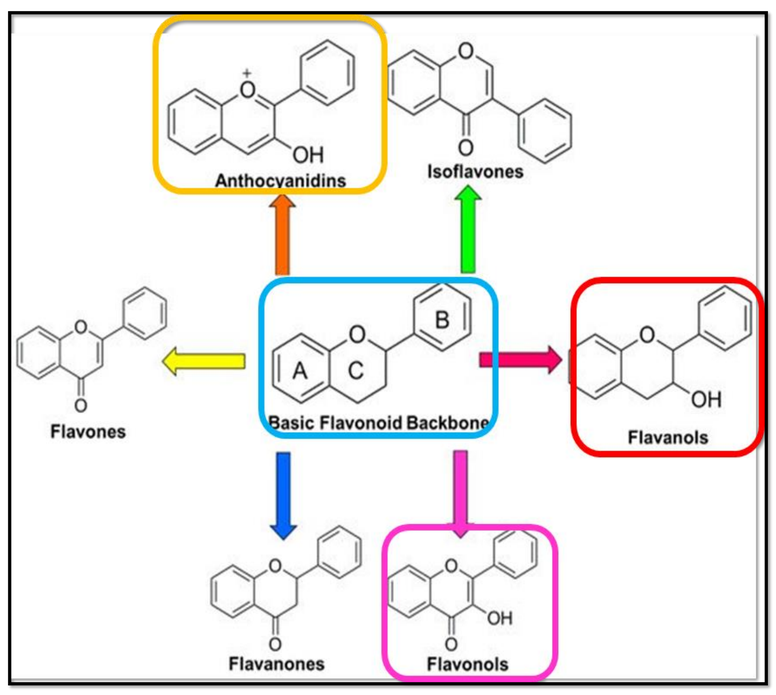
$$QC_1 = \sqrt{\frac{\sum_{i=1}^{N}((y_i - \hat{y}_i)/\hat{y}_i)^2}{N-1}} \qquad QC_3 = \sqrt{\frac{\sum_{i=1}^{N}((y_i - \hat{y}_i)/\bar{\hat{y}}_i)^2}{N-1}}$$

$$QC_5 = \sqrt{\sum_{i=1}^{N}(r_i/\max|r_i|)^2} \qquad \mathrm{NQC}_5 = \frac{QC_5 - 1}{\sqrt{N} - 1}$$

$$QC_6 = \sqrt{\sum_{i=1}^{N}(r_i/\bar{r})^2} \qquad \mathrm{NQC}_6 = \frac{QC_6 - \sqrt{N}}{N - \sqrt{N}}$$

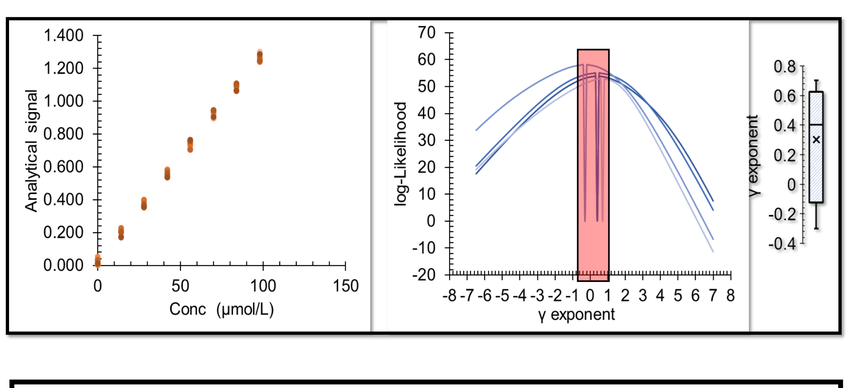
Evaluation of Gauss-Markov's premises for 11 experimentally

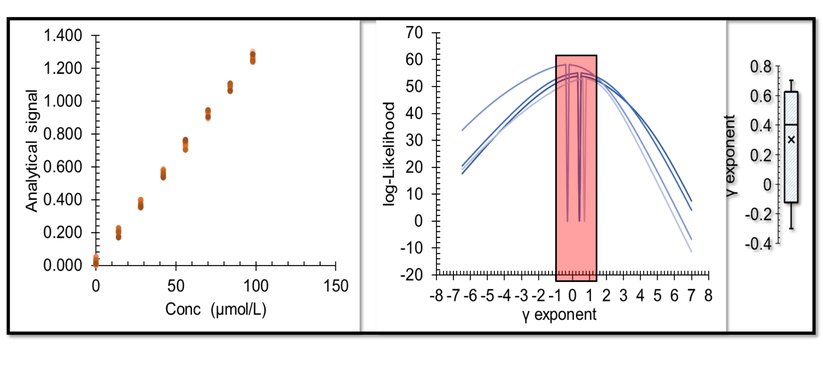


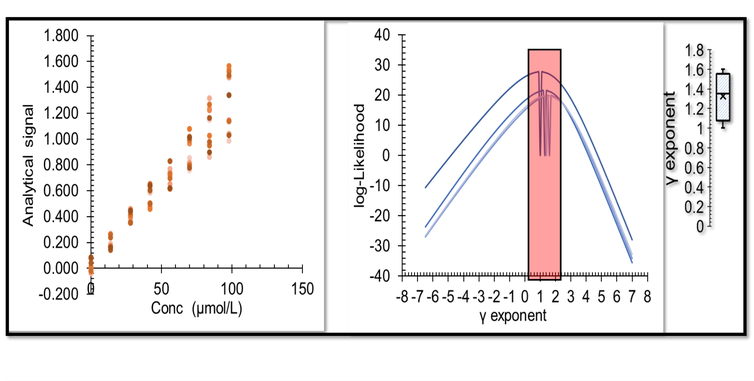


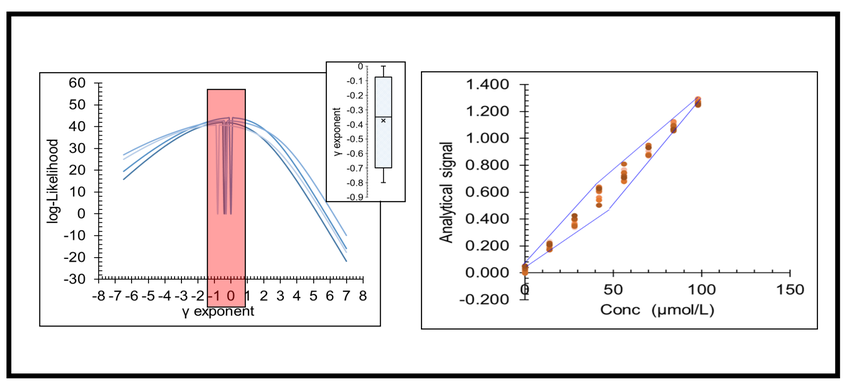
Evaluation of Gauss-Markov's premises for 11 experimentally obtained calibration data sets

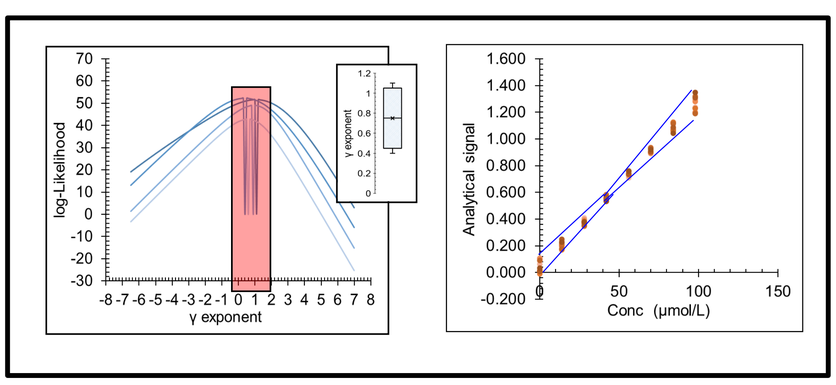
No.	Analytical method for total phytocompounds	Outliers present?		Test for heteroskedasticity		Linearity evaluation				
						Correlation coefficient		Mandel test for linearity		
		$\varepsilon_i > 2\sigma$	Cook's dist > 1	F-test	Breusch-Pagan	linear model	quadratic model	residual variance change (%)	F_{Mandel}	p-value
1	Total polyphenols FC	No	No	0.9160	0.9000	0.9980	0.9986	-9.2	2.2853	0.1920
2	Total polyphenols PB	No	No	0.1080	0.0850	0.9986	0.9993	-21	4.6950	0.0550
3	Total polyphenols FB	No	No	0.5220	0.4490	0.9977	0.9979	6.4	0.3031	0.9190
4	Total flavonoids M1	No	Yes (#8th)	0.0410	0.0400	0.9996	0.9999	-46	15.604	0.0042
5	Total flavonoids M2	No	No	0.2060	0.1570	0.9988	0.9988	7.1	0.2348	0.9467
6	Total flavonols	No	Yes (#8th)	0.0660	0.0560	0.9977	0.9994	-45	14.607	0.0049
7	Total flavanols	No	Yes (#8th)	0.0010	0.0080	0.9988	0.9989	3.0	0.6552	0.6913
8	Total o-diphenols	No	No	0.6070	0.5410	0.9994	0.9994	9.1	0.0385	0.9995
9	Total anthocyanins	No	No	0.0450	0.0420	0.9983	0.9984	6.6	0.2847	0.9210
10	Total coumarins	No	Yes (#8th)	0.2840	0.2210	0.9903	0.9988	-62	36.809	0.0006
11	Total alkaloids	No	Yes (#8th)	0.0210	0.0270	0.9994	0.9995	1.7	0.8019	0.6077

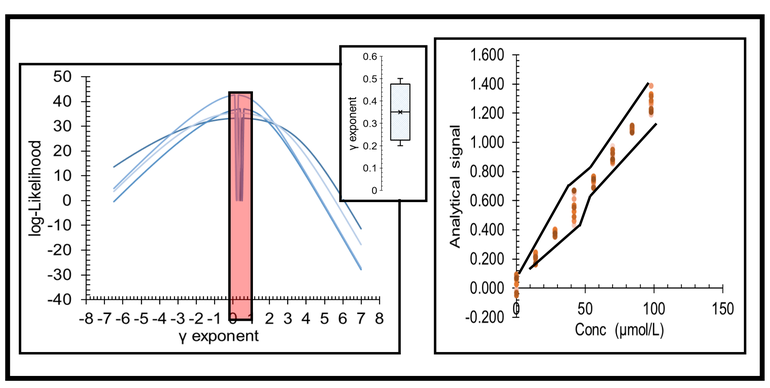












Conclusions

- **Gauss-Markov's premisses** (linearity, outlier, heteroscedasticity) for linear regression work successfully evaluated for **11 experimental data calibration sets** that were used in **total phytoconstituent content determinations** for several plant - based samples.
- A way of escaping the **heteroscedastic effect** is done by weighting with various **weighting factors**. Following the study, we propose an **optimal weighting factor** that would eliminate all difficult calculations and their time consuming.
- Regardless of the severity of the heteroscedasticity, the optimal range of the weighting factor will also take into account small dispersion values and high dispersion values. **The exponent gamma can be considered a measurement of heteroscedasticity** and it shows us how pronounced the heteroscedasticity phenomenon is. The **likelihood function** profile describes the heteroscedasticity behaviour.

Reference

[1] Sanchez M. J., Linear calibrations in chromatography: The incorrect use of ordinary least squares for determinations at low levels, and the need to redefine the limit of quantification with this regression model, *Journal of separation science*,**2020**, volume 43, pages 2708-2717

[2] Alladio E., Amante E., Bozzolino C., Seganti F., Salomone A., Vincenti M., Desharnais B., Experimental and statistical protocol for the effective validation of chromatographic analytical methods, *MethodsX*, **2020**, volume 7

[3] Özdemir Ş., Güney Y., Tuaç Y and Arlsan O., Empirical likelihood estimation for linear regression models with AR(p) error terms with numerical examples, *Journal of Applied Statistics*, **2021**, pages 1-16

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number TE-2019-1396, within PNCDI III