Linear Regression Modeling and Validation
Strategies for Structure-Activity Relationships

Sorana D. Bolboaca¹, Lorentz Jantschi²

¹ Department of Medical Informatics and Biostatistics, ‘Iuliu Hatieganu’ University of Medicine and Pharmacy Cluj-Napoca, 400349 Cluj-Napoca, Cluj, Romania.
sbolboaca@umfcluj.ro

² Department of Physics and Chemistry, Technical University of Cluj-Napoca, 101-103 Muncii Bvd., 400641 Cluj-Napoca, Cluj, Romania.
lorentz.jantschi@gmail.com

Keywords: quantitative structure-activity relationships (QSARs), linear regression analysis, model design, validation and diagnosis.

Identification and development of a new active compounds is an extremely expensive (reflected in time - between 10 and 15 years [1] and costs) and difficult process without a guaranteed result [2] (~ 90% of the initial candidates fail to be produces due to their toxicological properties [3]). Traditional strategies based on experiments (animal models [4]) are not anymore able to meet the actual needs in identification of new active compounds while in silico approaches such as computer-aided drug design [5], structure-based drug design [6], or virtual screening [7], are used nowadays.

Quantitative structure-activity relationships (QSARs) are mathematical relationships linking chemical structure and pharmacological activity/property in a quantitative manner for a series of compounds [8]. The approaches are based on the assumption that the structure of chemical compounds (such as geometric, topologic, steric, electronic properties, etc.) contains features responsible for its physical, chemical and biological properties [9]. The linear regression analysis is the statistical method frequently used in QSAR analysis since the main aim of the modeling is to identify a model able to predict the activity of new compounds [10].

Problems solving strategies in linear regression modeling include approaches for dealing with effective assessment of assumptions (linearity, independence of the errors, homoscedasticity, normality [11]), which seems to be broken in QSARs analyses [12,13]; effective methods for model selection [11,14,15]; efficient methods for model diagnosis [16,17]; and adequate approaches for assessment of predictive power of a QSAR model [18,19].

Here we emphasize problem solving strategies that address the main issues that arise when developing multivariate linear regression models using
real data.

Other problems not addressed here include the dealing with not normal distributed errors [20,21] and additional methods for estimation of regression parameters [22].

References


