MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **53** (2005) 175-180

ISSN 0340 - 6253

Residual Plots and the Quality of a Model

Lionello Pogliani*, Jesus V. de Julián-Ortiz+

Dipartimento di Chimica, Università della Calabria, 87030 Rende, Italia.
lionp@unical.it.
*On sabbatical leave, see next address.
+ Unidad de Investigación en Diseño de Farmacos y Conectividad Molecular, Facultad de Farmacia,
Dept. de Química Física, Av. V.Andrés. Estellés s/n, 46100 Burjassot (València), Spain.
julian@goya.combios.es

(Received July 23, 2004)

Abstract

Five residual plots of modelled properties, as well as a random residual plot are analyzed to underline the importance of residual plots in detecting anomalies in the quality of a model. Residual plots can help to detect anomalies, which hide inside tables of modelled properties or activities, and which are not easily detected by the statistical parameters. They also tell us that a greater care should be taken in the analysis of calculated data, which are not always as optimal as they seem to be from some validation methods.

Introduction

Recently [1] plot methods, whose importance was already underlined in a less recent review [2] have been reconsidered in the light of a series of publications centered around new considerations about statistical methods and analysis in QSAR/QSPR model studies. [3-9] Plot methods in QSAR/QSPR studies, continue to be considered as an optional by many authors and, in a less plausible way, by referees too. In fact, papers without any plot and, instead, full of Tables of observed vs. calculated values continue to appear with constant pace throughout the normal scientific literature. Furthermore, experimental vs. calculated plots accompanied with the corresponding residual plots are, practically, non-existent. Now, if till some years ago, due to the objective inherent limits of the used PCs, which normally did not have any graphical software, this could be an excusable matter, nowadays it is less and less excusable, as any PC is equipped with an easy to handle graphical software and, in some cases, with more than one graphical software. As already said [1, 2] plots methods can illustrate and detect violation of assumptions; that is, values should show random fluctuations around the main diagonal of the observed vs. calculated plot. That is equivalent to saying that residuals, the difference between experimental and predicted values, should show random fluctuation around a value of zero. Clusters of positive and negative values might suggest that a curvilinear trend in the data should be investigated. In a set of values obtained in sequence, there should not be long runs of values on the same side of the main diagonal of the figure; that is, there should not be systematic trends in the sequence of residuals, even if it is difficult to quantify what constitutes a long run. Employing plot methods it is easier to detect the presence of outliers in the data set, which lead to an inflated standard deviation and, in some cases, this allows a strategy to be outlined for their treatment.

To base a model on statistical parameters only can be rather misleading. To add to these parameters tables of observed plus calculated data, with the corresponding residuals and even left-out values, can even be more misleading as it may convey the impression to the non highly attentive reader that the authors have done the best they could. Now, to detect limits in a model by the help of statistical values and tables alone one needs a lot of attention and patience, i.e., definitely a lot of time, and time is not exactly what everyone has at disposition by full hands. What normally happens is that tables are read 'diagonally' and the statistical parameters end up assuming an overarching meaning.

In this paper we will center our attention on the residual plots, which in normal QSAR/QSPR literature are even rarer than the normal observed/calculated plots. Actually, they can be much more informative, and they can convey the limits of a model in a more striking way. Here, we are going to analyze six residual plots of six different properties, P_1-P_6, whose observed vs. calculated values have been taken from the current literature on the subject.

Discussion

It is highly advantageous for their analysis to divide the residual plots into at least three different regions around the null line, [1] i.e., the ± *s*, ± *2s*, and ± *3s* regions, where *s* is the standard deviation of estimates. Now, residuals around the null line should be completely random, i.e., no patterns should be detectable around this line. Furthermore all the points should, possibly, lie within the ± *3s* region. For possibly it is meant that some points could lie outside, but they should be the large minority. Within the ± *s* limit should possibly fall 60%-70% residuals, while within the ± *2s* range should possibly fall 90-95% residuals. Clearly, all this depends on the aims of the model. Sometimes even highly approximated models are of good help.

Throughout Figure 1 is shown the residual plot of $\mathbf{P_1}$ (left) and $\mathbf{P_2}$ (right) properties for forty-eight chemicals. At the top of this figure are shown the corresponding statistics.

$\mathbf{P_1}$ (left) F = 661, r = 0.967, s = 3.7, n = 48; $\mathbf{P_2}$: F = 1633, r = 0.986, s = 0.7, n = 48

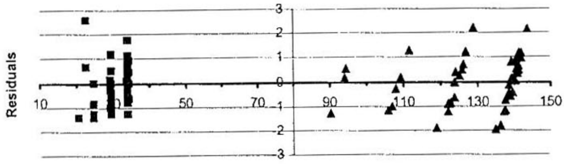

Figure 1. Residual plot for the P1 (*left*) and P_2 (*right*) properties.

The best statistics are those of $\mathbf{P_2}$, whose residual plot is shown in Figure 1-right. This figure is saying that (*i*) residuals are not random, (*ii*) data are clustered and (*iii*) the residuals are 'fanning out', i.e., they grow from left to right, from le left-most cluster to the right-most cluster. Each cluster seem, thus, to constitute by itself a new domain of data which should be modeled separately from the other clusters. The residual plot for $\mathbf{P_1}$, on the left, shows anomalies in the left-most cluster. The detection of clusters normally requires a deeper analysis of the data in order to gain a better insight about the model of the corresponding property. A small set of clusters are easier to model than a large number of non-clustered data, and could be welcome only if the residuals do not diverge within the cluster, and from cluster to cluster.

The plot of the next property, $\mathbf{P_3}$, is shown in Figure 2 (left), and its statistics are shown on the top at the figure. Here the residual plot does not allow for any doubt, deviations are not at all random and they follow an evident sigmoidal pattern, which goes unnoticed in the normal observed vs. calculated plot. Further, a good deal of residuals fall in the ± *3s* region, with a residual in the ± *4s* region. Probably, a simple linear regression equation is not sufficient for the present model. The residual plot of $\mathbf{P_4}$ is shown in Figure 2-rigth, and its statistics are shown at the top of the figure. This rather good residual plot shows nevertheless residuals that 'fan out' from left to right. The whole has the form of a triangle with the vertex on the left, and the basis on the right. The triangle base might be considered a cluster by its own. This unusual form points to a suspicious case of heteroscedasticity, which seems also to be shared by the $\mathbf{P_1}$ and $\mathbf{P_2}$ properties. Actually, heteroscedasticity means non-constancy of the variance of a measure over the levels of the property under study, to cope with some tests have been developed. [10] More data would be needed here to get rid of this suspicion.

P_3: F = 560, r = 0.95, s = 9.5, n = 62 ; P_4: F = 2615, r = 0.993, s = 5.8, n = 39

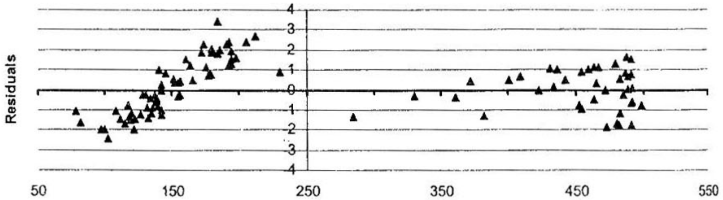

Figure 2.(left) Residual plot for P_3; *right*: residual plot for P_4 (this plot has been shifted by hundred units to avoid overlap with the left plot).

In fact, to check for homo- or heteroscedasticity the total number of observations should be divided into smaller and roughly equal subsets, and the variance recalculated for each subset. Clearly, if with more data the residuals would continue to 'fan out' the entire model could become unreliable, as the predictive ability for the lower values of P is different from the predictive ability for the upper values of P.

The next property to be considered is $\mathbf{P_5}$, whose residual plot is shown in Figure 3, and whose very satisfactory statistics are,

P_5: F = 1010, r = 0.991, s = 15, n = 40

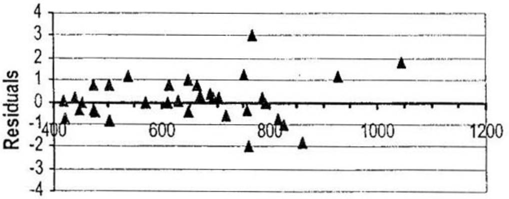

Figure 3. Residual plot for the P_5 property.

This plot is better than any of the previous residual plots. The numbers of points in the ± *2s* and ± *3s* regions is well within the optimal suggested percentages. The residual plot for $\mathbf{P_6}$ with only 20 points and a common statistics is shown in Figure 4,

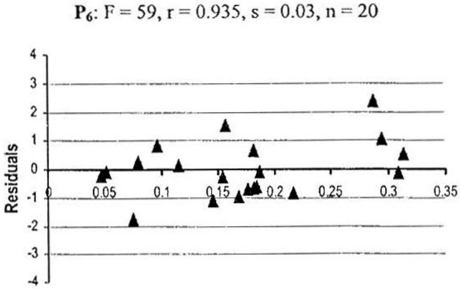

Figure 4. Residual plot for the P_6 property.

This residual plot, even if obtained with only twenty points shows no suspect flaws: nearly all points are within the ± *1s* region, few points are in the ± *2s* region, and only a point exceeds the ± *2s* region. Further, there is no clear evidence either of clustering or of heteroscedasticity.

For comparison purposes in Figure 5 (left and right) is shown the residual plot for two different sets of fifty evenly spaced Gaussian random numbers obtained with a procedure drawn in ref. [11] Normally, residuals should approximate this behavior.

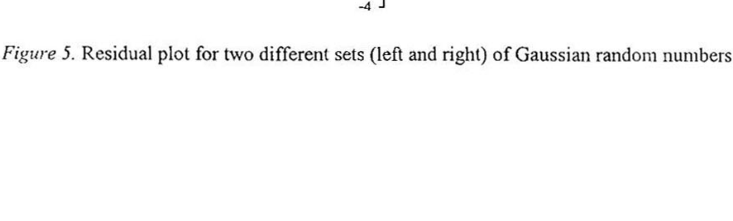

Figure 5. Residual plot for two different sets (left and right) of Gaussian random numbers.

Conclusions

The surprising 'anomalous' trends of the residual plots $\mathbf{P}_1$-$\mathbf{P}_4$ can, practically, be detected only with plot methods, and especially with residual plots, as abnormal trends are normally overlooked in tables of experimental vs. calculated values. To rely heavily only on statistical parameters can sometimes be misleading, as the first four residual plots show. The rather good residual plot of $\mathbf{P}_6$ underlines the fact that it is not always a question of having a small data set. Clearly, residual plots do not represent the decisive trademark for the overall quality of a model, nevertheless they constitute an important step, after the step of the statistical parameters, to ascertain that a model is on the right track, and possibly to correct it. Further, the higher sensibility of residual plots is a guarantee against flaws concealed in the observed vs. calculated plots. Clearly, the quality of a model depends on the aims of the model, and should possibly encompass other statistical methods to further check its value.

Acknowledgements

The first author would like to thanks Professors J. Galvez and R. Garcia-Domenech of the University of Valencia, Spain, for their and assistance. The second author would like to thank the '*Red de Investigación de Centros de Enfermedades Tropicales*', Ministry of Health, Spain.

References

[1] L. Pogliani, J.V. de Julian-Ortiz, Chem.Phys.Lett. 393 (2004) 327.

[2] L. Pogliani, Chem.Rev. 100 (2000) 3827.

[3] J. Pecka & R. Ponec, J.Math.Chem. 27 (2000) 13.

[4] A. Golbraikh & A. Tropsha, J.Molec.Graph.&Model. 20 (2002) 269.

[5] A. Golbraikh, M. Shen, Z. Xiao, Y. Xioa, K. -H Lee & A. Tropsha, J.Comput.-Aid.Mol. Des. 17 (2003) 241.

[6] D.M. Hawkins, S.C. Basak, & D. Mills, J.Chem.Inf. Comput.Sci. 43 (2003) 579.

[7] D.M. Hawkins, J.Chem.Inf.Comput.Sci. 44 (2004) 1.

[8] C. Peterangelo, & P.G. Seybold, Int.J.Quant. Chem. 96 (2004) 1.

[9] J.V. de Julián-Ortiz, E. Besalú & R. García-Domenech, Ind.J.Chem. 42A (2003) 1392.

[10] J.D. Lyon & C.L. Tsai, The Statistician 45 (1996) 337.

[11] A. R. Miller, Turbo Basic Programs for Scientists and Engineers (Sybex, New York, 1987) ch. 2.