International Journal of Pure and Applied Researches http://ijopaar.com; 2018 Vol. 1(1); pp. 110-121, ISSN: 2455-474X

Optimization of QSAR Models for Predicting Anticancer Activity of Gossypol Acetic Acid against the Anticancer Target BCL2 using Multiple Linear Regression and Feature Selection Techniques

Rajeev Pandey*

&

Varun Kumar Kashyap**

*Corresponding Author, Department of Statistics, University of Lucknow, India;

Email: profrajeevlu@gmail.com

**Assistant Professor, Department of Community Medicine, Jamia Hamdard Institute of Medical Sciences and Research, New Delhi, India.

Abstract

In recent years, statistical modelling techniques have gained significant traction; however, their application across various fields has been inadequate, often lagging behind. The field of drug design is no exception. This article presents the development of a statistical model based on an empirical dataset, specifically focusing on the exploration of the quantitative structure-activity relationship (QSAR) within an anticancer protein cell dataset. The study investigates the anticancer activity of Gossypol acetic acid against the BCL2 target, specifically for colorectal cancer, breast cancer, and mouth cancer. The model is developed using 80% of a virtual sample comprising 138 data points. The regression coefficient of Multiple Linear Regression (MLR) is computed for the training set, utilizing the Leave-One-Out (LOO) method for cross-validation. The remaining 20% of the dataset serves as a test set to validate the proposed model. Five influential factors with a high degree of statistical efficiency have been detected.

Keywords: QSAR, anticancer activity of Gossypol acetic acid, MLR and Leave-One-Out.

1. Introduction

There are an estimated 2.5 million cancer cases in India, with over 800,000 new cases and 556,400 recorded fatalities each year. Among the recorded deaths in a study of 122,429 individuals, 7,137 were attributed to cancer, reflecting the national estimate of 556,400 cancer-related deaths in India in 2010. A significant portion of these deaths, around 71%, occurred in individuals aged 30-69 years, accounting for 200,100 men and 195,300 women. Within this age group, the three most common fatal cancers in men were oral cancers (including lip and pharynx) with 45,800 cases (22.9% of deaths), stomach cancer with 25,200 cases (12.6% of deaths), and lung cancer (including trachea and larynx) with 22,900 cases (11.4% of deaths). Among women aged 30-69 years, the three leading causes of cancer-related deaths were cervical cancer with 33,400 cases (17.1% of deaths), stomach cancer with 27,500 cases (14.1% of deaths), and breast cancer with 19,900 cases (10.2% of deaths).

QSAR modelling, which stands for Quantitative Structure-Activity Relationship modelling, is a widely utilized approach in chemistry to predict and understand the properties and activities of chemical compounds. By analysing the relationship between the molecular structure of a compound

and its activity or property, QSAR models provide valuable insights for drug discovery, toxicity assessment, environmental impact analysis, and other chemical-related applications. The ability to predict chemical properties accurately can save time, resources, and effort in the development of new compounds or the evaluation of existing ones. The selection of appropriate statistical methods is crucial in QSAR model development. Different statistical approaches, such as linear and non-linear methods, have distinct characteristics and assumptions. The methodology selected can have a considerable impact on the model's precision, robustness, and forecasting abilities. Understanding the strengths and limitations of

different statistical methods allow researchers to make informed decisions in selecting the most suitable approach for their specific QSAR modelling goals.

MATERIALS AND METHOD

Regression analysis is considered to be one of the most widely used statistical techniques used for analysing multifactor data. Its broad appeal is a result of the conceptually simple process of using an equation which expresses a relationship between a set of variables. Successful regression analysis requires an appreciation of both the theory and pragmatic problems that often arise when the technique is used for the real-world data. The regression models are used for many purposes including data description, parameter estimation, prediction and control when a regression equation is used prediction purpose, it is important that the variables must be related in a casual manner.

1.1 Multiple Linear Regression (MLR)

In order to establish a relationship between **X** and **y**, Multiple Linear Regression (MLR) has until recently been the widely used method of choice. In MLR, it is assumed that **X** is of full rank and the *xij* are measured with negligible error. The algebraic MLR model is defined in Equation 1.1 and in matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (1.1)$$

where $\mathbf{X} = [\mathbf{x}_0|\mathbf{x}_1|\ldots\mathbf{x}_J]$, $\boldsymbol{\beta}' = [\beta_0, \beta_1,\ldots, \beta_J]$ and **e** is an error vector. Note that the first column in **X**, *i.e.*, $\mathbf{x}_0$ consists of only constants which, after mean-centering, becomes zero and consequently $\mathbf{x}_0$ is omitted. When **X** is of full rank the least squares solution is:

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (1.2)$$

where $\hat{\beta}$ is the estimator for the regression coefficients in β. An obvious disadvantage using MLR as regression method in QSAR is: when $I \leq J$ **X** is not of full rank and $(X'X)^{-1}$ in Equation 2.3.1, is not defined and β cannot be estimated. In this chapter we have also discussed about the problem with multicollinearity, *i.e.*, the case when **X** not is of full rank.

1.2 Multicollinearity

In the previous section, the potential danger of multicollinearity in combination with MLR was mentioned. Multicollinearity is present when the columns of **X** are approximately or exactly linearly

dependent. In the case of exact linear dependency, $(X'X)^{-1}$is not defined, and the estimation of the regression coefficients $\hat{\beta}$ cannot be expressed as in Equation 1.1 anymore. If the linear dependency is approximate, assuming **X** is properly auto-scaled, at least one of the diagonal elements in the inverse covariance matrix, $(X'X)^{-1}$, will be large. Additionally, some of the diagonal elements of cov$\hat{\beta}$, well-known to be s2$(X'X)^{-1}$ (where s2 is $\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\hat{\beta}\right)/(I-J)$ $(I > J)$), may be large, indicating that some b^s in $\hat{\beta}$ are estimated with low precision. Consequently, multicollinearity may influence the interpretation of the model and affect external predictions detrimentally. Therefore, it is important to be able to detect whether **X** is collinear or not, prior to regression analysis. The inverse covariance matrix, $(X'X)^{-1}$, provides a first indication of ill conditioning (multicollinearity) among the variables in **X**. Another commonly used indication of multicollinearity is the variance inflation factor (VIF):

$$VIF_i = 1/1 - R_i^2 \qquad (1.2.1)$$

where R_i^2 is the squared multiple correlation coefficient when **X***i* (the *i*th variable in **X**) is regressed on the remaining variables. When the columns of **X** are close to linear dependence (*i.e.*, when the determinant of $(X'X)^{-1}$ is close to zero), R_i^2 will be close to unity and VIF*i* will be large. In the ideal case, when $(X'X)^{-1} = I$, *i.e.*, when the variables in **X** are orthogonal, the VIF for the *i*th variable is unity. Thus, the VIF measures the increase (inflation) of the variance, for each variable, compared to the ideal case. A flag of warning is raised when VIF is greater than five, as suggested by **Smilde.**

The condition index or number (f) is defined as:

$$ø = \frac{\lambda_{max}^{0.5}}{\lambda_{min}^{0.5}} \qquad (1.2.2)$$

Where λ_{max} and λ_{min} represent the largest and the smallest eigenvalue, respectively, of $(X'X)^{-1}$ (scaled and centered **X**). When **X** is ill-conditioned, at least one eigenvalue will be close to zero and, consequently, f becomes large. As a rule of thumb, when f exceeds 100, the effect of multicollinearity may be significant. The influence of multicollinearity in QSAR is well known, and disqualified MLR as regression method years ago. In a chemical system, controlled by variables that are easily manipulated, an experimental design may be a solution to avoid multicollinearity. In QSAR, however, the objects are generally molecules which can complicate an experimental design.

1.3 THE *n* VARIABLE MODEL: NOATATION AND ASSUMPTIONS

Generalizing the n –variable population regression function (PRF).we may write the n-variable PRF as

$$Y_i = \beta_1 + \beta_2 X_{2_i} + \beta_3 X_{3_i} \ldots\ldots\ldots\ldots \beta_{ni} + \mu_i \qquad (1.3.1)$$

Where Y is the dependent variable, $X_1, X_2 \ldots\ldots\ldots\ldots\ldots\ldots X_n$ are the explanatory variables (or repressors), μ the stochastic disturbance term, and i the i[th] observation in case the data are time series the subscript t will denote the t[th] observation. In E_q (2.5.1) β_1 is the intercept term. Generally it gives the mean or average effect on Y of all the variables excluded from the model although its mechanical interpretation is the average value of Y when X_1, X_2……X_n are set equal to zero. The coefficients β_1, β_2, β_3, ……..β_n are called the **partial regression coefficients** and their meaning will be explained shortly.

Researcher operated within the framework of the classical linear regression model (**CLRM**) specifically researcher assumes the following:

Zero mean value of μ_i

$$E(\mu_i|X_1, X_2, \ldots X_n) = 0 \text{ for each } i \qquad (1.3.2)$$

No serial correlation or

$$cov(\mu_i, \mu_j) \forall\, i \neq j \qquad (1.3.3)$$

Homoscedasticity or

$$var(\mu_i) = \sigma^2 \forall i \qquad (1.3.4)$$

Zero covariance between μ_i and each X variable, or

$$cov(\mu_i, X_{2_i}) = cov(\mu_i, X_{3_i}) = \cdots = cov(\mu_i, X_{n_i}) = 0 \quad (1.3.5)$$

No specification bias, or the model is correctly specified.

No multicollinearity between the X variables, or No **exact linear relationship** between X_2, X_3, ……..X_n.

Informally no collinearity means of the regressors can be written as exact linear combinations of the remaining regressors in the model. Formally no collinearity means that here exists no set of numbers λ_1, λ_2, ……λ_n not both zero such that

$$\lambda_2 X_{2_i} + \lambda_3 X_{3_i} \ldots\ldots\ldots \lambda_n X_{n_i} = 0 \qquad (1.3.6)$$

If such an exact linear relationship exists then X_2, X_3,……..X_n are said to be **collinear** or linearly dependent .On the other hand. If true only when $\lambda_2 = \lambda_3 = 0$ thenX_2 and X_3 are said to be linearly independent.

Thus if

$$X_{2_i} = -4X_{3_i} or\ X_{2_i} = 4X_{3_i} = 0 \qquad (1.3.7)$$

The n variables are linearly dependent and provided both are included in a regression model we will have perfect collinearity or an exact linear relationship between the n regressors.

First the assumption of no multicollinearity pertains to our theoretical ($i.e$, PRF) model. In practice when we collect data for empirical analysis there is no guarantee that there will not be correlations among the regressors. As a matter of fact, in most applied work it is nearly impossible to find out two or more variables that may not be correlated to some extent.

Second, we are talking only about perfect linear relationships between two or more variables. Multicollinearity rule out nonlinear relationship between variables Suppose$X_{3_i} = X_2^2$. This dose not violate the assumption of no perfect collinearity as the relationship between the variables here is nonlinear.

1.4 OLS ESTIMATION OF THE REGRESSION COEFFICIENTS:

To obtain the OLS estimate of β, let us write k variable simple regression:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \dots\dots\dots\dots\dots\dots\dots\dots \hat{\beta}_k X_{ki} + \hat{u} \qquad (1.4.1)$$

This can be written more compactly in matrix notation as

$$y = X\hat{\beta} + \hat{u} \qquad (1.4.2)$$

and in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

Where $\hat{\beta}$ is a k-ellement column vector of OLS estimator of the regression coefficient and where $\hat{u}$ is a $(n \times 1)$column vector of n residual.

As in k – variable case the OLS estimator is obtained by minimizing

$$\hat{u} = y - X\hat{\beta} \qquad (1.4.3)$$

In OLS method RSS minimise with respect to parameter β the RSS is given by

$$\hat{u}'\hat{u} = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)$$

$$= yy' - 2\hat{\beta}X'y + \hat{\beta}'X'X\hat{\beta} \qquad (2.6.4)$$

Where use is made of the properties of the transpose of a matrix, namely, $\left(X\hat{\beta}\right)' = \hat{\beta}X'$

And since $\hat{\beta}X'y$ is a scalar (a real number), it is equal to its transpose $y'X\hat{\beta}$.

In scalar notation, the method of OLS consists in so estimating β_1, β_2, ……..β_k that $\sum \hat{u}^2{}_i$ is as small as possible. This is done by differentiating (eq.) partially with respect to

$$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \hat{\beta}_k$$

and setting the resulting expression to zero. This process yield k simultaneous in k unknowns, the normal equation of the least- square theory.

$$(XX')\hat{\beta} = X'y \qquad (1.4.5)$$

Note this feature of $(X'X)$ matrix:

(1) it gives the raw sum of square and cross products of the X variables, one of which is intercept term taking the value of 1 for each observation. The elements on the main diagonal give the raw sum of square, and those off the main diagonal give the raw sums of cross products (by raw we mean in original unit of measurement).

(2) It is symmetrical since the cross product between $X_{(k-1)i}$ and X_{ki} is same the same as the between X_{ki} and $X_{(k-1)i}$.

(3) It is of order (k x k), that is, k rows and k columns.

In (1.4.5) the known quantities are $(X'X)$ and $(X'y)$ (the cross product between the X variables and y) and the unknown is $\hat{\beta}$. Now using matrix algebra, if the inverse of $(X'X)$ exists, say$(X'X)^{-1}$, the pre multiplying both side of (1.4.5) by this inverse, we obtain

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y$$

But since $(X'X)^{-1}(X'X) = I$, an identity matrix of order k x k, we get

$$I\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (1.4.6)$$

Equation (1.4.6) is a fundamental result of the OLS theory in matrix notation. It shows how the $\hat{\beta}$ vector can be estimated from the given data.

Result and discussion:

2.1. Model development

A dataset comprising functional analogs centered around gossypol acetic acid with anti-BCL2 activity was initially collected from the NCBI database. Two-dimensional molecular descriptors were computed for each compound to digitize the observational data. A total of 255 descriptors were calculated using the PaDEL software developed by the National University of Singapore, providing a

comprehensive representation of the structural characteristics of the molecules. Initially, all 255 descriptors were computed for each compound. However, not all of these descriptors significantly contributed to the bioactivity. Therefore, several steps were taken to eliminate less informative descriptors, including:

1. Eliminating descriptors with constant values.
2. Removing descriptors with more than 90% zero values.
3. Excluding descriptors with constant or zero variance.

Following these steps, highly correlated descriptors were excluded using a correlation matrix approach. Descriptors with a correlation coefficient greater than 0.3 (positive or negative) with the bioactivity vector of the available datasets were retained. Consequently, only five descriptors remained for further analysis. This matrix-based feature reduction technique effectively reduced the variable space and minimized the chances of correlation between descriptors. The removal of correlated descriptors reduced noise in the data, resulting in a final dataset consisting of 106 compounds with their corresponding activity and five selected descriptors. The selected descriptors for modelling purposes were identified as MDEC.33, MDEO.11, MDEO.12, MDEO.22, and MLFER_S, which exhibited a strong association with activity and were deemed significant. For detailed descriptions of the descriptors used, further information can be accessed from the PaDEL descriptors website at http://www.ncbi.nlm.nih.gov. The coefficient matrix provided below highlights the variables that are considered significant in the model.

Variable	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-8.38500E+02	1.36800E+03	-6.13000E-01	0.5421
StsC	-6.73800E+00	3.37600E+00	-1.99600E+00	0.0504
SdssC	-2.07700E-01	3.15900E-01	-6.57000E-01	0.5133
SssssC	2.61500E+01	2.39400E+01	1.09300E+00	0.2788
SdsN	-7.10800E-01	3.77900E-01	-1.88100E+00	0.0647
SsssN	-1.68300E+00	1.20400E+00	-1.39700E+00	0.1672
SdO	1.52700E-01	1.16800E-01	1.30700E+00	0.196
SssO	3.24100E-01	3.23900E-01	1.00100E+00	0.3209
SsOm	9.90500E-03	3.03700E-01	3.30000E-02	0.9741
SsF	3.80100E-02	1.06100E-01	3.58000E-01	0.7213

SssS	5.81600E-02	7.12600E-01	8.20000E-02	0.9352
SddssS	-6.49000E-02	2.59500E-01	-2.50000E-01	0.8033
SsCl	5.61600E-01	1.82000E+00	3.09000E-01	0.7587
SsBr	-3.57200E+00	2.32100E+00	-1.53900E+00	0.1289
fragC	-7.56900E+00	4.23600E+00	-1.78700E+00	0.0789
nHBAcc	7.23800E+00	1.03900E+01	6.97000E-01	0.4887
Kier1	-1.05600E+03	9.48100E+02	-1.11400E+00	0.2697
Kier2	-4.69700E+01	1.86100E+02	-2.52000E-01	0.8015
Kier3	5.27400E+00	4.94900E+01	1.07000E-01	0.9155
nAtomLC	2.03200E+03	1.27500E+03	1.59400E+00	0.116
nAtomP	2.05200E-02	2.03600E-02	1.00800E+00	0.3174
nAtomLAC	-2.98500E+01	6.03600E+01	-4.95000E-01	0.6227
MLogP	1.50000E+00	1.02200E+01	1.47000E-01	0.8838
McGowan_Volume	-2.47000E+01	1.36600E+02	-1.81000E-01	0.8571
MDEC.11	-1.64600E+00	1.39000E+00	-1.18400E+00	0.2408
MDEC.12	-1.92500E-03	2.80000E-01	-7.00000E-03	0.9945
MDEC.13	4.61600E-02	1.71000E-01	2.70000E-01	0.7881
MDEC.14	4.67600E+01	4.28400E+01	1.09200E+00	0.2793
MDEC.22	-6.95400E-03	4.10900E-02	-1.69000E-01	0.8662
MDEC.23	1.37700E-02	1.87700E-02	7.34000E-01	0.4658
MDEC.33	-5.05200E-02	2.19400E-02	-2.30300E+00	**0.0247***

MDEO.11	-2.33500E+00	1.02700E+00	-2.27400E+00	**0.0264***
MDEO.12	-4.54800E+00	2.09000E+00	-2.17600E+00	**0.0334***
MDEO.22	2.12600E+01	8.08600E+00	2.62900E+00	**0.0108***
MDEN.11	-6.41300E-01	3.13400E+00	-2.05000E-01	0.8385
MDEN.12	7.89100E+00	1.12900E+01	6.99000E-01	0.4874
MDEN.13	-4.30800E+00	3.58700E+00	-1.20100E+00	0.2343
MDEN.22	4.81500E+00	5.52900E+00	8.71000E-01	0.3872
MDEN.23	1.91300E-01	1.87400E+00	1.02000E-01	0.919
MDEN.33	-5.54200E-01	3.57200E+00	-1.55000E-01	0.8772
MLFER_BH	6.38100E+01	3.86800E+01	1.65000E+00	0.1041
MLFER_BO	-6.08300E+01	3.82400E+01	-1.59100E+00	0.1168
MLFER_S	-4.56400E+00	2.24900E+00	-2.02900E+00	**0.0468***
MLFER_E	2.48700E+00	2.12000E+00	1.17300E+00	0.2452

QSAR model equation –

Predicted log IC_{50} (µM) = -0.05052 x (MDEC.33)

-2.33500 x (MDEO.11)

-04.54800 x (MDEO.12)

+21.26000 x (MDEO.22)

-4.56400 x (MLFER_S)

-838.50

[Regression coefficient (r^2) = 0.7314 and Cross validation coefficient (rCV^2) = 0.8299]

The developed QSAR model equation demonstrated a relationship between in vitro experimental activity (IC50) and five chemical descriptors. The model exhibited a regression coefficient (r2) of 0.73, indicating a 73% correlation between activity and descriptors in the training dataset. The cross-validation regression coefficient (rCV2) was 0.82, indicating an 82% prediction accuracy of the QSAR model. The analysis revealed that the descriptor MDEO.22 displayed a positive correlation

with increased biological activity against lung cancer cell lines, while MDEC.33, MDEO.11, MDEO.12, and MLFER_S showed negative correlations, indicating that higher values of these descriptors corresponded to decreased activity.

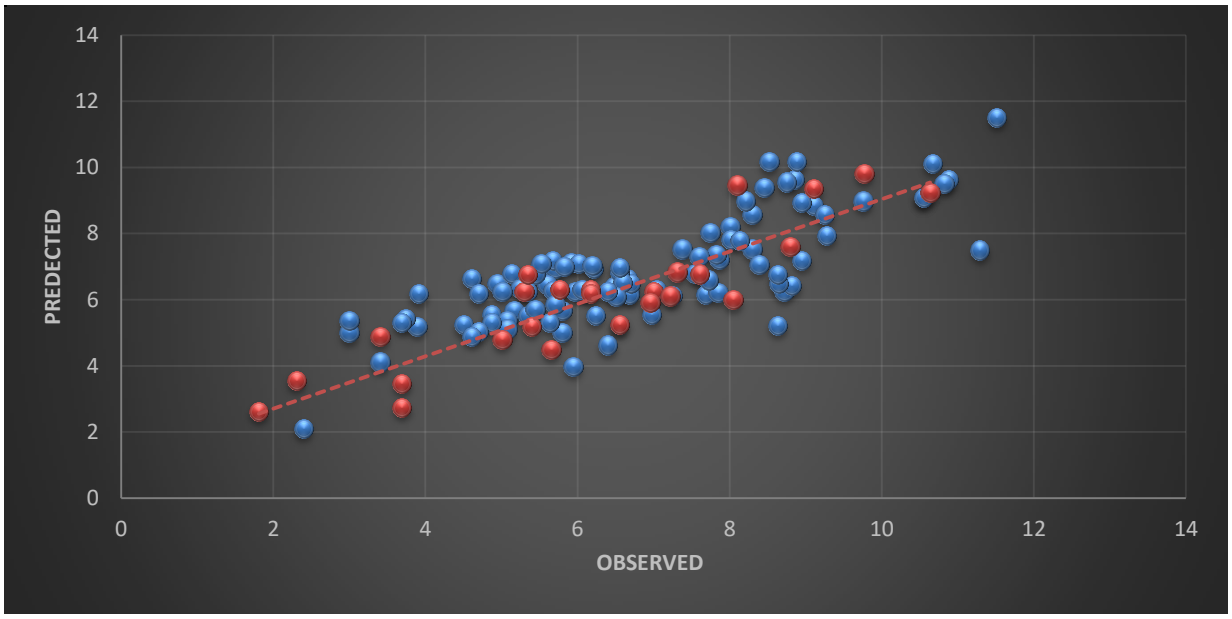

Figure 2.1. Graphical plot of multiple linear regression analysis which indicates linear relationship between experimental and predicted log IC_{50} with r^2= 0.73.

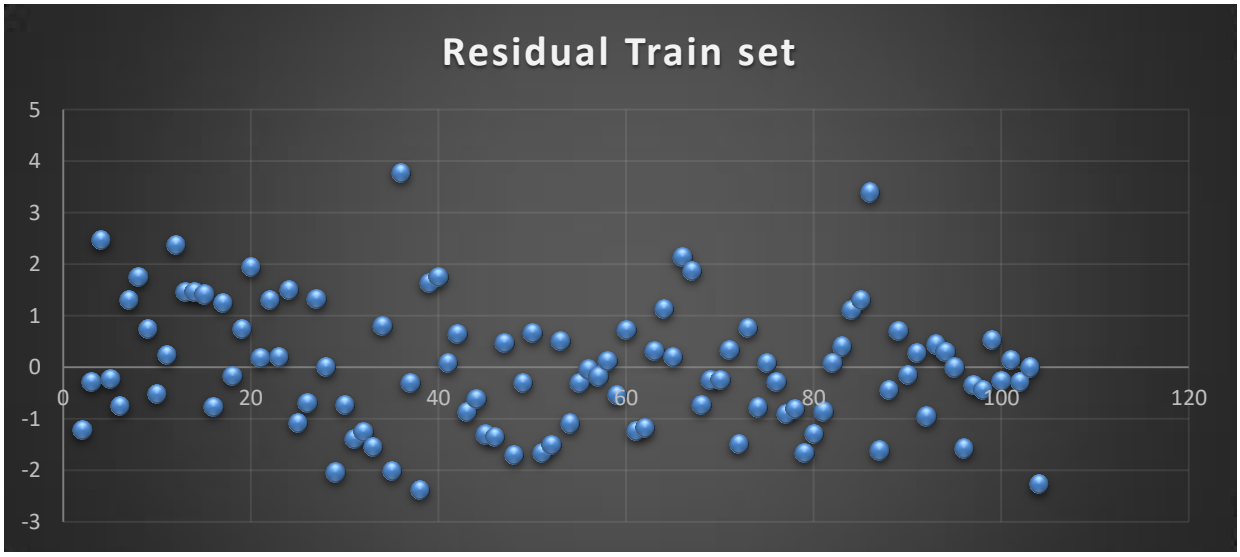

Figure 2.2. Residual plot of multiple linear regression analysis for experimental and predicted log IC_{50}.

2.2. Validation of QSAR model

To test the internal stability and predictive ability, QSAR model was validated by the internal, external validation and randomization test procedure:

2.3. Internal Validation

Internal validation was carried out using leave-one-out (LOO) method. For calculating cross validation regression coefficient (rCV^2), each molecule in the training set was eliminated once and the activity of the eliminated molecule was predicted by using the model developed by the remaining molecules. The cross-validation regression coefficient (rCV^2) was calculated using the equation which describes the internal stability of a model.

$$r^2 = 1 - \frac{\Sigma(Ypred - Y)^2}{\Sigma\left(Y - \dot{Y}\right)^2}$$

2.4. External Validation

For external validation, the activity of each molecule in the test set was predicted using the model developed by the training set. The regression coefficient (r^2) value is calculated as follows.

$$r^2_{cv} = 1 - \frac{\Sigma\left(Y_{pred(test)} - Y_{(test)}\right)^2}{\Sigma\left(Y_{(test)} - \dot{Y}_{(training)}\right)^2}$$

Thus, the regression coefficient (r^2) value is indicative of the predictive power of the current model for external test set. For this researcher has used only eight compounds for test. Generally, a QSAR model was considered to have a high predictive power only if the rCV^2 was >0.6 for the test set.

2.5. CONCLUSION

The robust QSAR model was developed by the multiple linear regression method in order to correlate the chemical structures of selected compounds to its reported anticancer activities. The correlation in terms of r^2 and prediction accuracy in terms of rCV^2 of derived QSAR model were 0.73 and 0.82 respectively. The QSAR study indicates that chemical properties *viz*., MDEC.33, MDEO.11, MDEO.12, MDEO.22and MLFER_S are correlate well with anticancer activity. These inferences and results will offer useful references to understand the molecular mechanism and to direct the design of new anticancer drug with improved activity.

References:

1. GLOBOCAN 2012: Estimated cancer incidence, mortality, and prevalence worldwide in 2012: Cancer Base No. 11 [Internet]. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C et al., editors. Lyon, France: International Agency for Research on Cancer; 2013. Available from: https://gco.iarc.fr/today/home
2. Balachandran S, Duraisamy S, Palanisamy S, Arun Kumar R, Nagarajan R. An overview of QSAR modeling: methods and applications in medicinal chemistry. Expert opinion on drug discovery. 2015 Jan 1;10(12):1281-97. DOI: 10.1517/17460441.2015.1073810

3. Roy K, Kar S, Das RN. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press; 2015. ISBN: 978-0-12-801773-5.
4. Katritzky AR, Lobanov VS, Karelson M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. 1995 Jan 1;24(4):279-87. DOI: 10.1039/CS9952400279
5. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research. Journal of chemical information and modeling. 2010 Nov 22;51(11):3143-6. DOI: 10.1021/ci2003869