

Regression Analysis Applied to Solubility Data: From Ordinary to Penalized Models

Brewton Morais
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
brewtonlmorais@gmail.com

Lucas Abdalah
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
lucasabdalah@alu.ufc.br

Abstract—Chemical and physical properties may be exploited to predict molecular interactions of solubility for drugs investigation. On this work we take advantage on mathematical tools, such as linear regression model analysis to perform prediction for data with information related to the chemical structure of compounds to predict their solubility. Based on data preprocessing techniques a data reduction approach is presented for the model. Linear relationship observed for some predictors is summarized with PCA and correlation matrix computing, to assess predictors relationship. In addition, a brief analysis of linear regression models and comparison to the penalized models, such as Lasso and Ridge. Finally, an analysis of the prediction accuracy of the developed model was performed with cross-validation techniques assessed by error parameters such as root mean squared error (RMSE) and coefficient of determination R^2 .

Index Terms—Chemical solubility, Linear regression, Machine learning, Penalized models, Principal component analysis.

I. INTRODUCTION

The solubility is defined by the International Union of Pure and Applied Chemistry (IUPAC) as "the analytical composition of a saturated solution, expressed in terms of the proportion of a designated solute in a designated solvent, is the solubility of that solute." [1], i.e., the physical property of substances to dissolve, or not, in a given liquid. The solubility may be expressed as a concentration, molality, mole fraction, mole ratio, etc.

This characteristic property of a specific solute-solvent combination plays a fundamental role in scientific research and practical applications, observed in medicines and drugs field, as an index to characterize chemical substances and compounds, and to assist in the identification of a compound polarity.

The popular Kosher or kitchen salt (NaCl) is formed by ionic bonds between Sodium (Na^+) and Chlorine (Cl^-) ions, resulting in a polar salt that is soluble in aqueous solution. This is due to its structure, composition, charge density, intermolecular attraction forces, carbon chain size (in the case of organic compounds).

For an analysis where the exact influence of each chemical and physical characteristic of the experiment compounds on the solubility of the product is not known, it is convenient to study the predictors and its impact to the output at each observation. However, different approaches may impose great

computational cost, in such way that a study exploiting all the data description available may keep redundant information, and lead to drawbacks in data representation and visualization. The main goal is to build framework capable of providing preprocessed data to use in a linear regression-based prediction model able to predict the data solubility. These steps precede more complex analysis such as feature extraction, dimensionality reduction, clustering and class separation [2]–[4].

The importance of these proposals is related to the applications, which can be in several areas, for example: with the financial market, economy in the large-scale use in the chemical industry, in the production in pharmaceutical laboratories, and even in the study of the influence of pollution on coral destruction.

Multiples approaches for the scenario are exploited in the literature with the same dataset or very similar, which linear regression models and neural networks are applied. Taking advantage on these mathematical tools, some papers explore the relationship of molecular topology predictors to infer the solubility of organic compounds in water [5], or in a similar approach, in addition to molecular topology, electronic interactions are explored in the so-called "E-state" [6]. For another dataset, yet working with the same chemical context and characteristics that propose to relate mineral solubility as a function of ionic strength and temperature [7].

In order to assess this method and overcome these limitations, we propose the application of statistics and linear algebra techniques to observe the relationship between predictors (chemical structure of compounds) and solubility. A model proposition that aims to obtain some pattern of behavior of this relationship, in order to support the prediction for unknown compounds and reduce complexity.

II. METHODS

A. Data Overview

For the development of the predictive models, we use the Solubility data set ¹ provided by the Springer book 'Applied Predictive Modeling' [3] available as a built-in package in the R Project for Statistical Computing [8].

¹<https://cran.r-project.org/web/packages/AppliedPredictiveModeling/>

The data consist of 1267 observations (samples) of various chemical compounds, with the solubility of the compound as the output. While the predictors are divided into three groups, the first with 208 binary predictors referring to the presence of a specific chemical substructure, the second with 16 predictors counting chemical bonds and certain types of atoms, and the third with four continuous predictors with characteristics of the compounds, such as molecular mass.

The binary predictors are labeled with *FP001* until the predictor *FP208*. The remaining 16 continuous predictors are identified as: *MolWeight*, *NumAtoms*, *NumNonHAtoms*, *NumBonds*, *NumNonHBonds*, *NumMultBonds*, *NumRotBonds*, *NumDblBonds*, *NumAromaticBonds*, *NumHydrogen*, *NumCarbon*, *NumNitrogen*, *NumOxygen*, *NumSulfur*, *NumChlorine*, *NumHalogen*, *NumRings*, *HydrophilicFactor*, *SurfaceArea1*, *SurfaceArea2*, *solubility*.

B. Notation

To ease the comprehension of this work, this section summarizes the notation used in the present paper and introduces some definitions.

Scalars, vectors, matrices are represented by lower-case (a, b, \dots), boldface lower-case ($\mathbf{a}, \mathbf{b}, \dots$) and boldface capital ($\mathbf{A}, \mathbf{B}, \dots$), respectively. The matrix transpose operator is represented by $(\cdot)^T$ and the symbol $(\hat{\cdot})$ represents an estimated value.

C. Data Preprocessing

Continuous and binary predictors contribute in different ways to the construction of the model, to define the best approach for each one.

For continuous data, we assess data mean, variance and skewness, to verify the existence of a significant presence of a left or right bias. In order to mitigate scale distortion between continuous predictors and others associated issues, we normalize the data before generating inferences from the data. We choose the z-score normalization, to perform a centering and scaling of the data to have a fair comparison of the histograms, correlation plots and further operators.

For normal distributions, only the knowledge of two statistics is enough to describe the set: mean (μ) and standard deviation (σ) of the samples, summarized as $\mathcal{N}(\mu, \sigma^2)$. The z-score, or standard score, is a technique that shift and rescale the data, providing a simplified version of each predictor representation, centered on zero, i.e., zero-mean distribution, and unitary standard deviation. So for a normal population distribution, it is enough to describe as $\mathcal{N}(0, 1)$.

We compute the sample mean (\bar{x}) for each predictor:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (1)$$

From the Eq. 1, we may obtain the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}, \quad (2)$$

Finally, from results of both equations 1 and 2, we obtain the transformed data z_i :

$$z_i = \frac{x_i - \bar{x}}{\sigma}. \quad (3)$$

A model constructed from a non-normalized data set may present a biased outcome, since the model is sensitive to variance. Moreover, in a non-normalized data, for large variance difference between the predictors, those with the greatest values misrepresent its relevance and distort the model, once it can be interpreted as delivering more relevant information, e.g., signal power/energy in digital signal processing field.

From the normalization we may use the covariance matrix and principal component analysis (PCA) to identify the predictors and its linear combinations that actually contribute to the model, producing a correlation analysis. It aims to identify those that carry redundant information, and in some cases eliminate it from the assessment. It also supports the identification of the best set of regression tools applied, because in case that too many predictors have non-linear or more complex relationship, more powerful tools may be applied due their capacity of dealing with such complicated problems.

It was checked if there were any missing values, but it was found that all data is fully completed, which eliminates the need of any data compensation technique.

D. Linear Regression

Linear regression is a simple and useful tool for supervised learning. It consists in an approach for predicting quantitative outcome \mathbf{y} based on a single predictor \mathbf{x} . It is assumed a linear relationship between predictor and outcome [4].

$$\mathbf{y} \approx \beta_0 + \beta_1 \mathbf{x}. \quad (4)$$

The model in Eq. 4 presents two constants, β_0 and β_1 , which stand for intercept and slope, respectively. Hence, the main goal is to estimate both coefficients to estimate an correlation between \mathbf{x} and \mathbf{y} .

We use a set for training to fit our estimated parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, to predict any sort of model based on linear correlation with a mean-zero random error term ϵ , a catch-all variable to accumulate what the model misses, since in general its true relationship is not linear [4].

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x} + \epsilon, \quad (5)$$

Taking advantage on the estimated parameters, we compute Eq. 5 to predict a continuous outcome.

However, the coefficients are unknown in practice and the objective of the ordinary least squares (OLS) linear regression is to find a plane that minimizes the residual sum of squares (RSS) between the observed data and the predicted response, i.e., the variance (σ^2) of the error.

$$\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2, \quad (6)$$

where the i -th term of e represents $y_i - \hat{y}_i$.

To assess the linear regression quality we may use the following indices: Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2 Statistic.

$$\text{MSE} = \frac{1}{n} \text{RSS} \quad (7)$$

The error variance usually is unknown, so we may estimate it from the data taking advantage on the model in Eq. 7. The idea is extended by computing the square root of this model to obtain the RMSE, i.e, $\text{RMSE} = \sqrt{\text{MSE}}$. Nevertheless, the value os these indices range accordingly with the y unit [3]. In order to overcome this limitation, the R^2 statistic provides another meausure of quality between 0 and 1, independent of y scale. In general terms, it provides an index to measure the amount of variability that is left unexplained in the model fit. It implies that as the R^2 values approximates to 1, a large proportion of the data variability is explained by the regression.

To compute the β parameters with a matrix approach, we may apply algebraic manipulation, assuming a non-singular matrix $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_n^T)$, where each \mathbf{x}_i is a predictor.

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (8)$$

The model shown in eq. 8 is an extension [3] to the concept presented in eq. 4, which is applied in this paper.

E. Penalized Models

The OLS regression approach provides an unbiased and low variance model. Although, this simple moodel present quite accurate predictions for proper data, its MSE perfomance can be improved by the addition of the sum of the squared regression parameters weighted by a penalization/regularization term (λ) [3].

$$\text{RSS}_{L_2} = \text{RSS} + \lambda \sum_{j=1}^P \beta_j^2 \quad (9)$$

The goal with the model presented in Eq. 9 is to allow a small increase in bias, which results in a substancial drop in the error variance. It imposes a new constraint to observe, the experimental search for an optimal λ value, to obtain an overall MSE lower than unbiased model [3], [4].

F. Principal Component Regression

Since the data set used has a large number of predictors, it would be interesting to reduce this number to make the model simpler and less computationally expensive. number of predictors, it would be interesting to reduce this number to make the model simpler and less computationally expensive. For this, one of the strategies is to find the so-called principal components, which are defined by the eigenvectors of the covariance matrix of the predictors. Thus, the data is projected onto a reduced number of predictors, those with the highest eigenvalues, that is, those with the greatest variability. There are two problems with this method, the first is that it becomes difficult to interpret the components, the second is that the method does not define the components by their relationship

with the output, which may cause the dominant components do not present a correlation with the output, which hinders the development of an efficient model, because you can not have control over the relationship of new predictors with the output.

G. Partial Least Squares

The partial least squares regression model can be seen as merging the features of linear regression models, which seek to maximize the correlation of predictors with output, and those in principal components, which capture the largest variances in the predictors. Thus, it is a supervised method that generates new components that have maximum covariance with the output, thus allowing a smaller number of components needed compared to PCR. However the problem of interpretability of the new predictors still remains. There are some algorithms for calculating PLS, such as NIPALS and SIMPLS.

H. Cross Validation

Cross-validation is a technique to evaluate the model within the test set, usually taking into account model flexibility and root mean square error. In short, the set is divided into k distinct groups of similar size, one of these groups is removed and becomes the validation set, then the models are produced from the remaining samples, and to check how well they work, the validation set is predicted. Then the removed group is returned to the training set, and the next group is removed and becomes the validation set. This process is repeated until all groups are used as validators.

This strategy often serves to indicate which models will predict better on the test set, as it allows the levels of error and variance generated by the models to be compared. Importantly, using k too small (as in only two groups $k = 2$) or too large ($k =$ number of samples) will create problems. In the first case there may be a high bias in the models, since many samples will be left out of the training group, and in the second case there will be too much variance in the models because the groups used in the method are very similar. Therefore, to compensate for both effects, it is usual to use $k = 5$ or 10 , since experimentally they present acceptable levels of variance and bias.

III. RESULTS

A. Covariance Matrix and Predictors Relationship

The degree of the linear relationship between two variables may be measured using the covariance coefficients (p). It ranges from -1 to +1, where large positive and negative values indicates positively and negatively correlated data, respectively. Its absolute magnitude measures the degree of redundancy. If the covariance is close to zero, the data is uncorrelated [3].

In order to provide data covariance, we compute the covariance matrix and present it as a heatmap for visualization simplicity. Fig. 1 shows that various predictors are strongly correlated, mainly positive.

Linear Regression Model
 Penalized Ridge Model
 Principal Component Regression
 Partial Least Squares

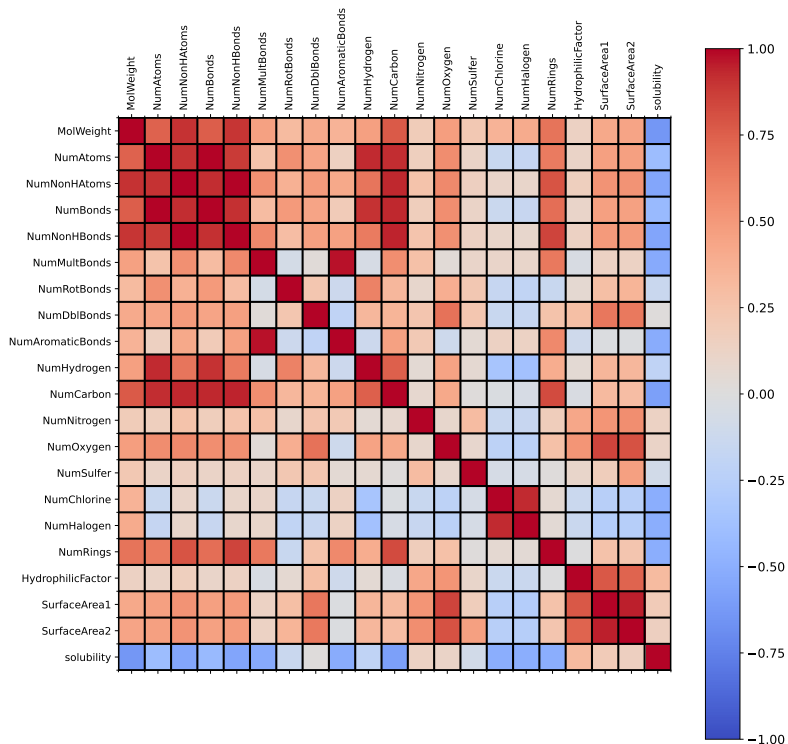


Fig. 1: Correlation Matrix presented as a Heatmap.

TABLE I: Covariance threshold analysis.

$ p \geq$	Related Pairs
0.75	25
0.90	14

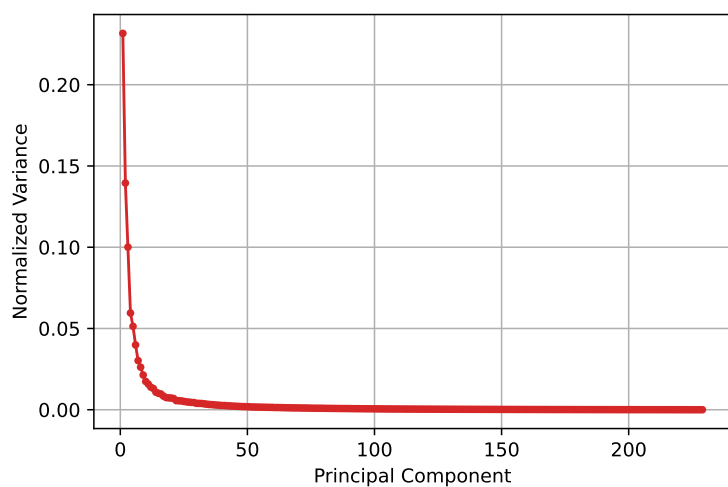


Fig. 2: Screeplot.

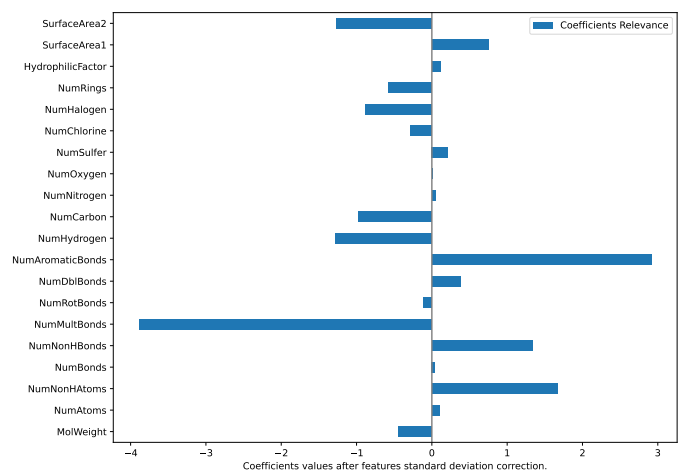


Fig. 3: Linear regression coefficients.

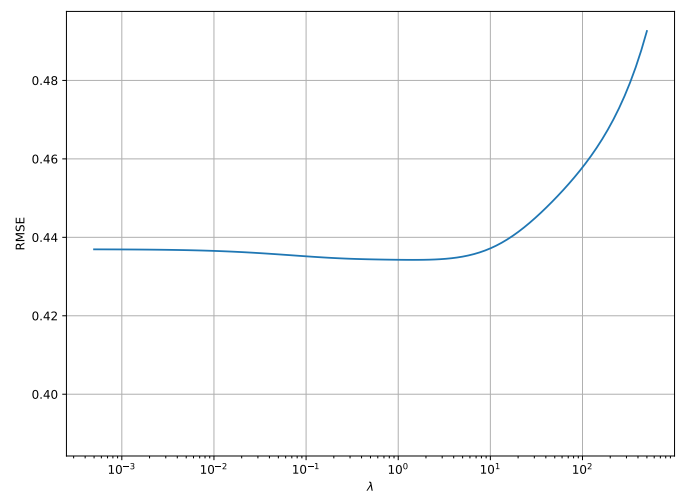


Fig. 4: λ Parameter vs. RMSE for a Ridge Model.

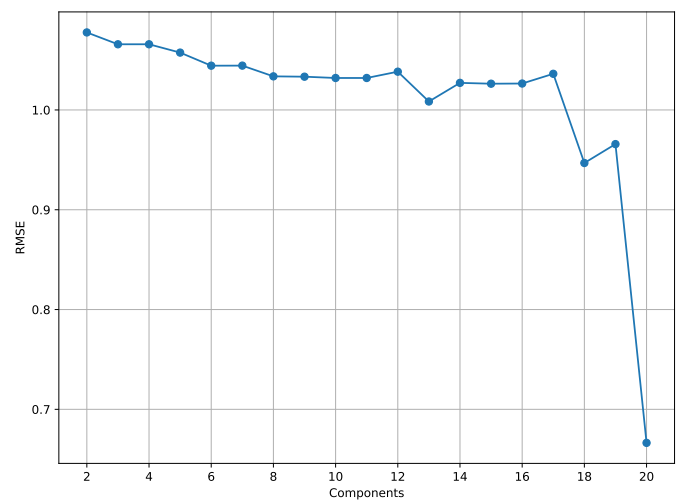


Fig. 5: Components in the Model vs. RMSE for a PCR Model.

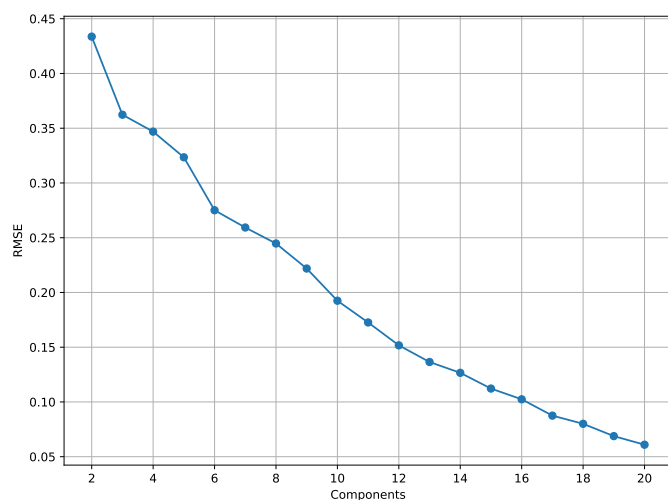


Fig. 6: Components in the Model vs. RMSE for a PLS Model.

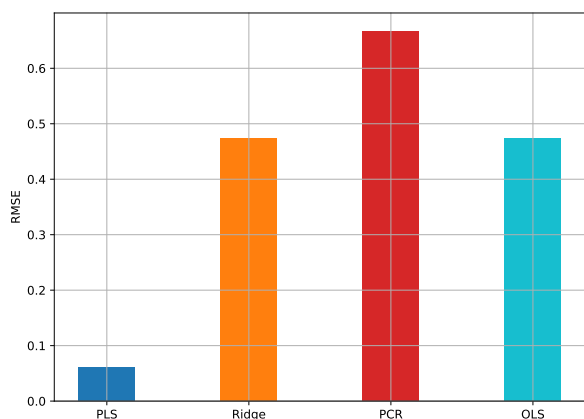


Fig. 7: Best value for RMSE for the proposed approaches.

IV. DISCUSSION

V. CONCLUSION

Among the results obtained, there were no great differences between the results of the models, which favors the use of simpler models and lower computational cost, to the detriment of little improvement in prediction with the increase of cost. Therefore, unless there is a real need to obtain the best possible results, the most cost-effective solution was simple linear regression, since it is inexpensive and obtained results that were not much inferior to its peers.

Nonetheless if computational cost is not taking in account, the best model for the given data set is the ridge regression model, since it presented the best results both in validation and in the test set.

VI. FURTHER WORK

APPENDIX

REFERENCES

- [1] A. D. McNaught and A. Wilkinson, *Compendium of Chemical Terminology: "The Gold Book"*. Oxford, 1997.
- [2] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.
- [3] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.
- [4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. [Online]. Available: <https://faculty.marshall.usc.edu/gareth-james/ISL/>
- [5] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, and A. E. P. Villa, "Estimation of aqueous solubility of chemical compounds using e-state indices," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 1488–1493, 11 2001.
- [6] J. Huuskonen, "Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology," *Journal of Chemical Information and Computer Sciences*, vol. 40, pp. 773–777, 05 2000.
- [7] S. M. Javed, "A regression model for mineral solubility as a function of ionic strength and temperature," Master's thesis, The University of Arizona, United States, 1991.
- [8] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: <https://www.R-project.org/>

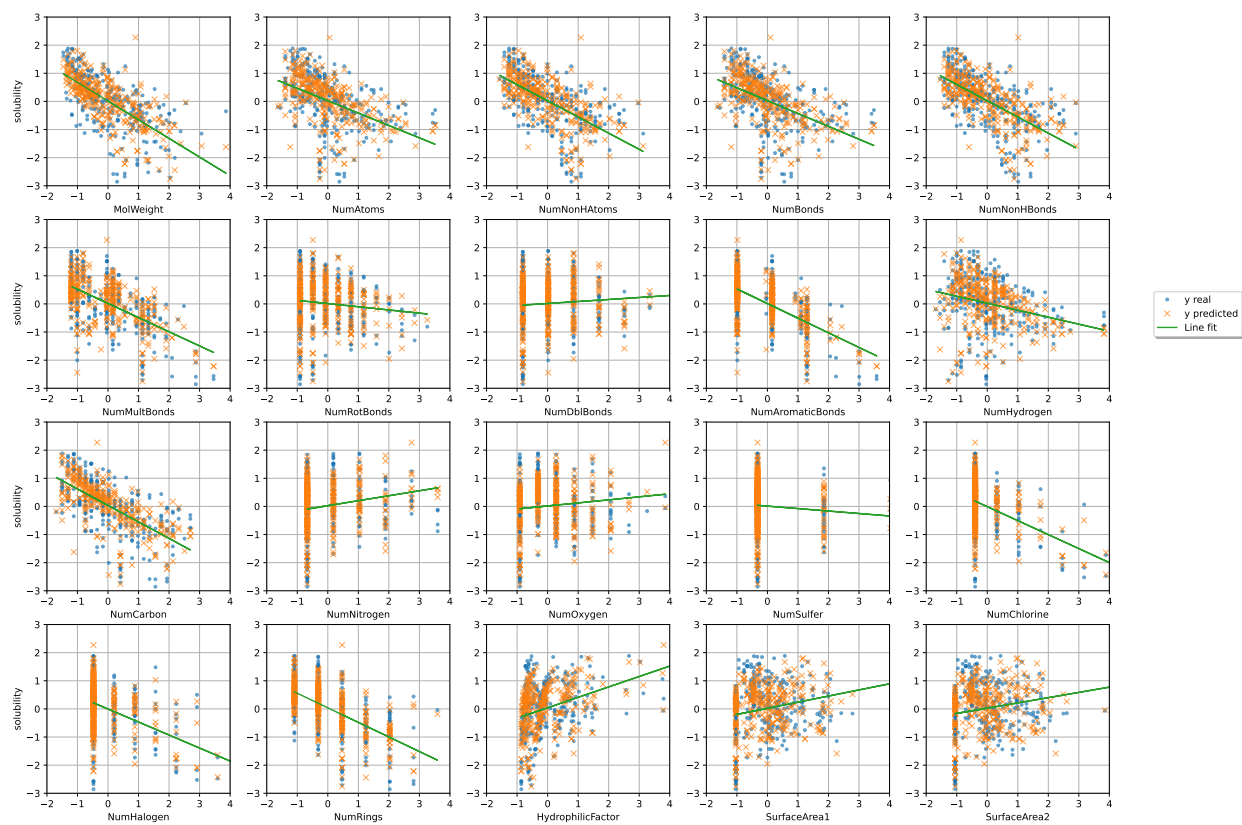


Fig. 8: Linear regression.