Appropriate Evaluation Measurements for Regression Models

Tsuyoshi Esaki1*

1The Center for Data Science Education and Research, Shiga University, 1-1-1 Banba, Hikone, Shiga 522-8522, Japan

*E-mail: tsuyoshi-esaki@biwako.shiga-u.ac.jp,

(Received July 5, 2021: accepted August 31, 2021; published online September 15, 2021)

Abstract

In recent years, accelerating the speed of finding seed compounds and reducing the cost of pharmaceutical research has become a necessity. The contribution of *in silico* drug discovery methods, which predict candidates as new drugs using physicochemical features and substructure fingerprints of compounds, is thus expected. Selecting the seed compounds without conducting experiments could enable us to reduce the time and cost required for drug development. However, estimating the characteristics of compounds in our body using a simple linear model alone is unsatisfactory because effects and distribution of compounds are determined by the environment in our body and their interactions with other molecules. Compared to simple models, more complex models have been prepared to estimate compound characteristics with high predictive accuracy. Thus, it is increasingly important to correctly evaluate the predictive performance when selecting the models appropriate for research purposes. The determinant coefficient, famous as $R^2$, is one of the most famous statistical measures for evaluating regression models. However, this measure cannot be used to evaluate nonlinear models. In this paper, the difficulty of using the determinant coefficient is explained and the proper statistical measures were suggested under the following two conditions: mean squared error (MSE) for cross-validation, and MSE along with correlation coefficients for the observed and predicted values of test data. As understanding statistical measures and using them appropriately is necessary, the suggested measures will support the effective selection of promising seed compounds and accelerate drug discovery.

Key Words: Regression model, *in silico* prediction, Data science, Model Evaluation

Area of Interest: *In silico* drug discovery, data science
1. Introduction

In recent years, accelerating the speed of identifying seed compounds and reducing the cost of pharmaceutical research have become a necessity. In silico drug discovery methods, which predict new drug candidates using physicochemical features and substructure fingerprints of compounds, are thus expected to contribute to this process. Furthermore, considering the difficulties in obtaining large amounts of experimental data with respect to animal safety, prediction algorithms that use artificial intelligence (AI) including machine learning methods are a promising approach. It is thus necessary to find seed compounds based on their chemical structure alone, without conducting any experiments or organic synthesis. Successful selection of seed compounds without experiments, could thus reduce the time and cost of drug development.

Prediction of drug efficacy and distribution is rather challenging. The lipoid theory of narcosis, a simple linear regression model for estimating the effect of narcotic drugs using an oil/water partition coefficient, suggested by Overton and Mayer from the end of the 19th to the early 20th century, was the first report addressing this aspect. Linear regression models were then used to predict aqueous solubility based on compound substructure until the year 2000. However, it is difficult to estimate the characteristics of compounds in our body using a simple linear model alone because their effects and distribution are determined by the environment in our body and their interactions with other molecules. Thus, nonlinear regression models constructed by machine learning algorithms were used after the year 2000, and since winning in the Kaggle competition in 2010, deep learning models have gained major attention [1, 2]. Especially, in recent years, descriptor-free prediction methods, such as graph convolution of deep learning, have been frequently used in in silico predictions of drug discovery [3]. This method does not require descriptor normalization or zero-value processing, and only requires the structure of compounds and labels; thus, it has an advantage of being easy to implement.

Compared to simple models, more complex models have been prepared to estimate the characteristics of compounds with high predictive accuracy. Developments in computer and programming libraries have enabled the construction of complex models with good performance. Thus, it is increasingly important to correctly evaluate the predictive performance when selecting the appropriate models for research purposes.

The determinant coefficient, written as $R^2$, is one of the most famous statistical measures for evaluating linear regression models. This measure cannot be used for evaluating nonlinear models, as reported previously [4]. The determinant coefficient has been used to evaluate the goodness of fit (GoF) of linear models in statistical research [5]. However, it has also been used to evaluate the predictive performance of nonlinear regression models in pharmaceutical or biochemical research fields [6]. This is because the major descriptions of regression model evaluation involve the determinant coefficient, whereas there are minimal descriptions regarding evaluation of prediction models in statistics textbooks. Statistical methods thus aim to examine “the reliability of coefficients fitted by having data,” whereas “the accuracy of predicted values against new data” is the most important aspect from a practical viewpoint. Many pharmaceutical researchers are thus interested in machine learning for the latter point.

Thus, an understanding of the methods used for the evaluation of regression models is required to select the appropriate statistical measures. In this paper, the appropriate statistical measures for practical machine learning projects under two conditions are suggested as follows: the mean squared error (MSE) for cross-validation, and MSE along with the correlation coefficients for the observed and predicted values in test data.
2. Model evaluation

Regardless of linear or nonlinear models, regression models are used for predicting values. Regression models are widely used to estimate the experimental values measured in drug discovery and safety experiments in the pharmaceutical field. Although various algorithms have been developed to construct regression models, such as multiple linear regression, support vector machine (SVM) with radial basis function (RBF) kernel, random forest (RF), and deep learning (DL), finding a good model is difficult because the appropriate algorithms are decided based on the collected dataset and research purpose. Thus, it is necessary to evaluate the predictive capability objectively and to separate “goodness of fit” and “predictive accuracy” [7].

2.1. Goodness of Fit (GoF)

Goodness of fit (GoF) provides a statistical measure for the data points used in the fitting parameters. The determinant coefficient (the defined equation is shown in Section 5) is used to measure the GoF of models. If this measure is high, the estimated model is generally good. However, this is not always the case. As an extreme example, the determinant coefficient of the fitted model is 1 when the number of descriptors is \((1 - \text{sample size})\), regardless of the randomized descriptors. This example indicates that excessive descriptors make the determinant coefficient an unsuitable measure for model fitting. Thus, an adjusted determinant coefficient (shown as adjusted \(R^2\)) was developed to avoid this problem. The adjusted \(R^2\) is decreased in value by the number of descriptors, similar to a penalty. Therefore, this measure is used to compare the proper number of descriptors for the model. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also used to select the proper descriptors and their numbers. The former measures the fitness of the model and descriptors under an estimated probability density. The latter measure calculates the fitness for the idea of Bayesian estimation. These measures indicate the fitness of having data points for the estimated models.

2.2. Predictive accuracy

Predictive accuracy (predictive performance) is a measure that shows how a model can predict new data correctly. The GoF shows the explainability of the data used for training. On the contrary, predictive accuracy shows the predictive capability of data that were not used for training. Thus, in cross-validation, which is used for parameter selection in machine learning, this measure is used for evaluation. Predictive accuracy is only calculated against the observed and predicted values, and the number of descriptors is not considered. The mean squared error (MSE) thus is the most useful measure for calculating predictive accuracy.

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

where \(y_i\) and \(\hat{y}_i\) \((i = 1, 2, ..., n)\) are the observed and predicted values, respectively. MSE is expected for the predictive accuracy of RF regression [8], which is an effective measure for both linear and nonlinear regression models. MSE is a measure used to calculate the variances between the observed and predicted values in the test data. Thus, it does not show a trend similar to the relationships in a scatter plot. Therefore, calculating the correlation coefficient between the observed and predicted values is also effective in comparing the ability as a relationship.
3. Linear and nonlinear regression models

A regression model is constructed to predict data or explain a phenomenon. The model combines the obtained phenomenon (outcome) and several factors (inputs or descriptors), considering the effect of the phenomenon. The regression model can then be separated into two models based on features such as the linear regression and nonlinear regression models.

In a linear regression model, the assumption is that the following function approximates the relationships between the inputs $x_{i1}, x_{i2}, \ldots, x_{ip}$ and the outcome $y_i$ for $i = 1, 2, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where the inputs are descriptors, and $\beta_j$ ($j = 1, 2, \ldots, p$) are the coefficients. The regression model is generally difficult for estimating the outcome perfectly because the outcomes are obtained from various experimental instruments or sensors, which are likely to generate noise related to the environment or experiment. The noise is shown as $\epsilon_i$ ($i = 1, 2, \ldots, n$) following a standard deviation. The coefficients of this model quantitatively explain the effect of the input on the outcome, which can provide helpful hints to improve pharmaceutical characteristics by modifying the descriptors of the compounds.

If target outcomes are difficult to predict using a linear regression model, complex models such as a nonlinear regression model can be useful. Several well-known nonlinear algorithms have been used in machine learning including SVM with RBF kernel, RF, neural network (NN), and DL. The characteristics of drugs are generally decided by the chemical substructure and physicochemical features, as well as by the environment in the body and interactions with other molecules. Therefore, nonlinear models have been reported to show a higher predictive performance compared to linear models. On the contrary, nonlinear models tend to have complex construction. Therefore, it is difficult to qualitatively and quantitatively explain the relationship between the inputs and outcomes. This means that it is difficult for nonlinear models to expound on themselves and to obtain hints for improving the structure. However, nonlinear models have been constructed to generate high-performance models because they can find seeds from thousands of compounds in the process of drug discovery.

The differences in the statistical measures of the models used are confirmed below.

4. Dataset

In this study, published data were employed to confirm the statistical measures in terms of the differences between linear and nonlinear regression models. The dataset contained the aqueous solubility measurements of 3,663 compounds [9]. Furthermore, 1,521 descriptors calculated using Dragon software [10] for each compound were also attached. All processes using this dataset were demonstrated using Python ver. 3.8. [11]

The dataset was randomly separated into a training set (70% of the dataset, 2,564 compounds) and test set (30%, 1,099 compounds) to analyze the data (Figure 1). The descriptors with low variance in the training set were removed using the VarianceThreshold function (threshold = 0.5) in the scikit-learn library (ver. 0.24.2) [12], because low-variance descriptors can cause overfitting. The number of retained descriptors was 428. Using the least absolute shrinkage and selection operator (LASSO) under 5-fold cross-validation, the proper penalty was determined as alpha = 0.00694. [13] This resulted in the retention of 131 descriptors, which were then used in further research.
5. Difficulty of R²

The determinant coefficient is a measure used to explain the GoF of a linear regression model, and this measure is often written as R². However, there are different expressions for R² in the literature. Eight patterns of expression have been described [4], and the more well-known expressions are as follows: (1) the squared multiple correlation coefficient between the regressand and regressor, (2) the determinant coefficient, and (3) the squared correlation coefficient between the observed and predicted values.

\[
R^2 = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n}(f(x_i) - f(x_i))^2}}
\]

Eq. (1)

\[
R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\]

Eq. (2)

\[
R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}}
\]

Eq. (3)

where \( f(x_i) \) \((i = 1, 2, ..., n)\) is a multiple linear regression function in the data points of \( x_i \). \( y_i \) and \( \hat{y}_i \) \((i = 1, 2, ..., n)\) are the observed and predicted values, respectively. \( \bar{y}_i \), \( \bar{y}_i \), and \( \bar{f}(x_i) \) are the mean values for \( y_i \), \( \hat{y}_i \), and \( f(x_i) \), respectively. Eq. (1) is a catchall term for the other equations. Suppose the number of descriptors is 1, R² becomes the Pearson product-moment correlation coefficient for the input and outcome in the bivariate linear regression model. Eq. (2) shows the variance ratio of the fitness because the determinant coefficient is calculated by the sum of the variances (residuals) between the observed and predicted values (left in Figure 2) divided by the sum of the variances of the observed values (right in Figure 2). Therefore, this measure is not valid if employed and compared for the same dataset because the average \( y \) value would be different.
Figure 2. Scheme by which variances are used to obtain the determinant coefficient

The determinant coefficient is obtained as the sum of the residuals between the observed values and fitted values (left) divided by the sum of the distance between the observed values and the mean of the dependent variable (right).

Eq. (3) shows the squared correlation coefficients. Forming a scheme for this is easy compared to that for the other equations.

Although the previous three equations have different definitions, they are often written as a unified term, \( R^2 \). The reason for this confusion may arise from the following theorem: “The model is a linear one and the ordinary least squares regression method is used for estimating the parameters. In this case, the previous equations are equal” [4]. From this theorem, the residual concerned with the prediction of a linear regression is explained by the sums of the total, residuals, and regression residuals (Total Sum of Squares = Sum of Squared Residuals + Regression Sum of Squares). If the model is nonlinear, this equation does not hold.

For this reason, the three equations with different definitions have been called the same name, \( R^2 \). This also results in confusion in the use of \( R^2 \).

5.1. Difference in \( R^2 \) between linear and nonlinear models

In the image of the differences between Eqs. (2) and (3), a multiple linear regression model was first constructed using the training data (left on Figure. 3). The trained model was then adapted for the test data (right on Figure. 3). The statistical scores are shown in Table 1. The result of Eq. (2) was calculated using the r2_score function of scikit-learn because scikit-learn learns Eq. (2) as \( R^2 \). [14]
Table 1. Summary of statistical scores for the results of linear and nonlinear models

| Algorithms                        | Training  | Test       |
|-----------------------------------|-----------|------------|
|                                   | Eq. (2)   | Eq. (3)    | Eq. (2)    | Eq. (3)    |
| Multiple linear regression model  | 0.8692    | 0.8692     | 0.7678     | 0.7757692  |
| (Linear model)                    |           |            |            |            |
| Neural network model              | 0.8820    | 0.8822     | 0.8129     | 0.81519    |
| (Nonlinear model)                 |           |            |            |            |

The training data were trained using the NN model with hidden_layer_size = 7 (Figure 4, left). The predicted result is shown on the right in Figure 4 to compare the linear model with the nonlinear model. A summary of statistical scores is shown in Table 1 with the results of the linear model.

Figure 4. Scatter plots of the fitted (left) and predicted (right) results
The x-axis and y-axis show the observed and predicted values, respectively. The black line represents the line at which the observed and predicted values are equal.

From Table 1, Eqs. (2) and (3) are equal only in the case of model fitting using the linear model. In other words, equality was not observed in the case of training with a nonlinear model or in evaluating the predictive accuracy of the test data.

5.2. Difference in $R^2$ between datasets

The determinant coefficient (Eq. (2)) used to evaluate the GoF of the linear regression model represents the variance of values estimated using the fitted model. However, the squared correlation coefficients between the observed and predicted values (Eq. (3)) show these relationships. To confirm the differences between the variances of the prediction, pseudo data were generated based on the test data. The data were random noise with different variances added as follows: variance = 0.05, 0.1, and 0.15 (Figure 5). The plots show that high variance data represented bad prediction results. A summary of the statistical scores is shown in Table 2. The correlation coefficient and MSE were calculated for each dataset to confirm the relationships and variances of the data. Although it was not appropriate to calculate the prediction accuracy, the determinant coefficient was also calculated for comparison with the other measures in this summary.
Figure 5. Scatter plots of the fitted result (x-axis) and predicted result (y-axis) at different variances
The x-axis and y-axis show the original values (pseudo-obtained values) and pseudo values (pseudo-predicted values), respectively. The black line represents the line at which the observed and predicted values are equal.

Table 2. Summary of statistical scores for the results of various data

| Variance     | Eq. (2) | Eq. (3) | Correlation coefficient | MSE     |
|--------------|---------|---------|-------------------------|---------|
| 0.05 (left in Figure 5) | 0.8931  | 0.8977  | 0.9475                  | 0.4623  |
| 0.10 (center in Figure 5) | 0.6297  | 0.6625  | 0.8139                  | 1.9308  |
| 0.15 (right in Figure 5)  | 0.3616  | 0.4561  | 0.6753                  | 4.1131  |

In this case, the determinant coefficient (Eq. (2)) was also not equal to the squared correlation coefficient between the observed and predicted values (Eq. (3)). Based on these results, the measure calculated with Eq. (2) is less than that calculated with Eq. (3); thus, the result in Eq. (2) can be a negative value because the lowest possible value of the result of Eq. (3) is 0. In fact, Eq. (2) can be a negative value because of its definition.

However, if considering the previous theorem that explains the equation of Eqs. (2) and (3), this is inconsistent and causes confusion regarding the understanding of $R^2$. Eq. (3) can be negative when the parameters are not appropriately estimated, or when outliers are included in the training set. This is also mentioned on the website of scikit-learn as “R² score may be negative” [14].

When the sum of the residuals is larger than the sum of the distances between the mean and fitted values, the determinant coefficient can be negative. This suggests that the model is not well defined (Figure 6).
5.3. Difference in $R^2$ between different software

$R^2$ has several definitions, and its calculated values are not always the same. Therefore, it is necessary to understand the mean $R^2$, which is now used appropriately. Different software use $R^2$ with different definitions, as follows: the determinant coefficient (Eq. (2)) in R and the Excel Linest function, the squared correlation coefficient between the observed and predicted values (Eq. (3)) in OpenOffice and PAST, and the other definition in Excel and GraphPad Prism [15]. Thus, the meanings of $R^2$ differ with the software used and it is not realistic to compare the $R^2$ values calculated using different software.

6. Suggested evaluation measures

The difficulty in using $R^2$ has been explained in the previous sections. This difficulty is possibly caused by limitations in usage of the determinant coefficient, the most used measure of $R^2$, which evaluates only the GoF of a linear regression model.

However, many researchers who use machine learning to predict their problems use cross-validation to select the appropriate parameters, and evaluate their model using test data. The appropriate evaluation in both cross-validation and evaluation with test data is predictive accuracy and not GoF. Furthermore, they often use a nonlinear regression model because of the necessity for a high predictive performance. This indicates that GoF is not necessary for evaluating models in general situations using machine learning.

Then, which evaluation measure is the most appropriate? The following measures are more appropriate for the cross-validation and evaluation of test data following the Model evaluation section (section 2) and references [7] and [8]:

- Cross-validation: MSE (or Root mean square error (RMSE))
- Evaluation of test data: MSE (or RMSE) and the correlation coefficient between the observed and predicted values with an explicit definition (or its squared value)

MSE is a useful measure for both linear and nonlinear models. Furthermore, it can be used for evaluating both cross-validation and test data, which is helpful to find overfitting by comparing with cross-validation and the results of test data (this is a piece of additional information; many packages for deep learning already employ MSE as the default for parameter searches). Further, it is easy to compare predictive accuracy with that of published models because several studies have used RMSE.

MSE shows the variation between observed and predicted values. Therefore, it does not represent the relationship between these, similar to the correlation coefficient. Thus, it is also useful to employ the correlation coefficient between observed and predicted values with an explicit definition (or its
squared value) to evaluate the predictive accuracy of test data. Using these two statistical measures enables confirmation of overfitting and evaluation of distances and correlation coefficients between the observed and predicted values.

7. Conclusion

Recently, the environment for easy construction of machine learning models has been improved by the development of various packages for demonstrating machine learning. However, there are problems in constructing prediction models with inadequate understanding, which can generate different results or unreliable decisions.

Some machine learning packages or libraries employ the determinant coefficient as the default for cross-validation despite the parameter fitting of a nonlinear regression model, which is a serious problem. Additionally, statistical measures being calculated despite inadequate evaluation is a major problem.

One of the reasons for this is that statistics textbooks mention the evaluation of predictive performance only briefly. This is because statistical importance and practical importance have different points of view. In practical machine learning problems, the required measures are MSE and the correlation coefficients between the observed and predicted values.

It is thus necessary for researchers to understand statistical measures and to use them appropriately. The suggestions in this paper will support the finding of promising seed compounds effectively and help accelerate drug discovery.

References

[1] Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.*, **2015**, *55*, 263–274. doi: 10.1021/ci500747n

[2] Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; *et al.* Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.*, **2017**, *57*, 2068–2076. doi: 10.1021/acs.jcim.7b00146

[3] Torng, W.; Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.*, **2019**, *59*, 4131–4149. doi: org/10.1021/acs.jcim.9b00628

[4] Kvålseth, T. O. Cautionary Note about $R^2$. *The American Statistician*, **1985**, *39*, 279–285. doi: 10.1080/00031305.1985.10479448

[5] Larry, W. Linear Regression and Logistic Regression. In *All of Statistics: A Concise Course in Statistical Inference*; George, C.; Stephen, F.; Ingram, O. Ed.; Springer, New York, 2013

[6] Spiess, A. N.; Neumeyer, N. An Evaluation of $R^2$ as an Inadequate Measure for Nonlinear Models in Pharmacological and Biochemical Research: a Monte Carlo Approach. *BMC Pharmacol.*, **2010**, *10*, 6. doi: 10.1186/1471-2210-10-6

[7] Galit, S.; Peter, C. B.; Nitin R. P. Evaluating Predictive Performance. In *Data mining for Business Analytics* (3rd ed); Wiley, New York, 2016, pp107

[8] Random Forests for Regression and Classification, https://math.usu.edu/adele/RandomForests/Ovronnaz.pdf (accessed Jun 1, 2021)

[9] Pirashvili, M.; Steinberg, L.; Belchi Guillamon F.; Niranjan, M.; Frey, J. G.; *et al.* Improved understanding of aqueous solubility modeling through topological data analysis. *J. Cheminform.*, **2018**, *10*, 54 doi: 10.1186/s13321-018-0308-5
[10] Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: an Easy Approach to Molecular Descriptor Calculations. *Match Commun. Math. Comput. Chem* 2006, 56, 237–248.

[11] Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace. Scotts Valley, CA, 2009 (accessed Jun 1, 2021)

[12] Pedregosa F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; *et al*. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830. doi:10.5555/1953048.2078195

[13] Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B* 1996, 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[14] R2_score, scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed Jun 1, 2021)

[15] Mistake of the determinant coefficient, https://biolab.sakura.ne.jp/r2.html (accessed Jun 1, 2021)