Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 10, Issue 1, pp: (20-24), Month: January - April 2023, Available at: www.noveltyjournals.com

A Comparative, Data-Based Assessment of Several Performance Metrics for Regression

[1]Jude Chukwura Obi, [2]Obimuanya Ijeoma

[1]Department of Statistics, Chukwuemeka Odimegwu Ojukwu University, Anambra State, Nigeria

[2]Department of Statistics, Chukwuemeka Odimegwu Ojukwu University, Anambra State, Nigeria

DOI: https://doi.org/10.5281/zenodo.7715854

Published Date: 10-February-2023

Abstract: **We have comparatively assessed five regression performance metrics namely, Mean Absolute Error, Mean Square Error, Root Mean Square Error, R^2 and $adjusted\ R^2$. A total of twelve datasets were used for the study inclusive of nine real-world datasets and three simulated datasets. Apart from the R^2 and $Adjusted\ R^2$, analysis shows that the output of other performance metrics are not data dependent. Regarding the simulated datasets which complied with all the regression model assumptions, we discovered that any of the five performance metrics can be used on the data, to assess the strength or otherwise of a regression model. This is based on the fact that the individual outputs of the performance metrics on simulated data are identical.**

Keywords: **Regression, Multiple Regression, Performance Metrices for Regression, Machine Learning.**

1. INTRODUCTION

Regression is an aspect of machine learning that focuses on the prediction of continuous outcome variables (Beers, 2003). As a machine learning tool, a regression model is trained and tested severally until an optimum model is obtained. Via the process of training and testing that is severally sustained, the regression model learns about the task it is expected to perform and given an out of sample data, such model usually performs optimally. A regression model can be simple or multiple, but the focus here is on the multiple regression model given in (1.1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \qquad (1.1)$$

One concern about the model specified in (1.1) is that in order to use it for prediction, it must be estimated. Since the model will be estimated, there are certain degrees of error that are evidently associated with the estimation. The smaller this error is, the better, meaning that the performance of any regression model is judged by the magnitude of error that associates with the model's estimation.

There are different ways of measuring the error resulting from estimating a regression model, and they altogether constitute what is called performance metrics for regression (Raj, 2022). They are termed performance metrics because they tend to show how a given regression model has performed. Essentially, these performance metrics include the following:

(a) Mean Absolute Error (MAE)

(b) Mean Square Error (MSE)

(c) Root Mean Square Error (RMSE)

(d) R Squared and

(e) Adjusted R Squared

Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 10, Issue 1, pp: (20-24), Month: January - April 2023, Available at: www.noveltyjournals.com

1.1 Mean Absolute Error (MAE)

Mean Absolute Error is the average of the absolute difference between the observed or given values and the predicted values. It gives the measure of how far the predictions are from the given values. Symbolically, MAE is as specified in (1.2).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{1.2}$$

1.2 Mean Squared Error (MSE)

The mean square error and mean absolute error are similar in appearance, but the obvious difference is that MSE is concerned with the square of the difference between observed and predicted values. For the reason that the square of the errors is taken, the effects of larger errors are more pronounced than smaller ones and it goes further to influence that magnitude of the MSE in the end. Symbolically, MSE is presented in (1.3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.3}$$

1.3 Root Mean Square Error (RMSE)

The root mean square error is simply the square root of the MSE. It is often preferred to the MSE because it is measured in the same units as the response variable (Bobbitt, 2021). The interpretation is more straightforward comparatively. For instance, A MSE with the value 16, equals a RMSE with the value 4. Now, a root mean squared error of 4 tells us that the average deviation between the predicted values and observed values is 4, as against the statement that the squared deviation (MSE) is 16. The RMSE is represented as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{1.4}$$

1.4 R-Squared

The R-Squared value known as coefficient of determination is a statistical performance measure for a regression model. It explains the proportion of variance for a dependent variable (Y) with respect to an independent variable (X) in the regression model. It is symbolically represented as given in (1.5).

$$\begin{aligned} R^2 &= 1 - \frac{SS_{Reg}}{SS_{Total}} \\ &= 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \end{aligned} \tag{1.5}$$

R-squared value lies between 0 and 1 inclusively, and should be as high as possible for the underlying model to be regarded as performing optimally. A lower value for R-squared points to the fact that the model's performance is not good enough.

1.5 Adjusted R Squared

The adjusted R-Squared (see eqn. (1.6)) computes the explanatory power of regression models that contain different numbers of predictors. It is a modified version of R-Squared that has been adjusted for the number of predictors in the model. Adjusted R Squared value considers only those independent variables which actually have effects on the performance of the model. A different use for the adjusted R-Squared is that it provides an unbiased estimate of the population R-squared.

$$Adjusted\ R^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{(n - p - 1)}\right] \tag{1.6}$$

2. AIM AND OBJECTIVES

The aim of this research is to compare the different methods of evaluating the performances of a regression model. The objectives will include the following:

- To find out if the performance metrics behave similarly on data or otherwise.
- To find out if based on the given dataset, a certain performance metric may be suggested for use.
- To show how the predicted response variable $(\hat{y})$ plays a crucial role in assessing the performance of a regression model.

In fact, it would appear that all the performance metrics focus on the prediction of the response variable $(\hat{y})$. The closer $\hat{y}$ is to Y, a given performance metric has a minimal error of estimation associated with it and vice versa.

3. RESEARCH METHODOLOGY

The research method here will involve the evaluation of a given regression model, on a variety of datasets using all the five metrics discussed in section 1.1. Regarding MAE, MSE and RMSE, the metric that will bring about the smallest error of estimation based on the given model, will be adjudge a better one. On the other hand, R^2 and $adjusted\ R^2$ recognize that assuming a perfect estimation of the response variable (Y) is achieved, both R^2 and $adjusted\ R^2$ will take the value 1. Now, a departure from 1 will help to determine the strength or otherwise of the regression model. For this reason, we would like to find out if a model that gives the lowest RMSE, for instance, is able to give highest score for R^2.

In fact, since we have several datasets, we would like to know if the performance metrics similarly behave given a dataset or whether their performances is data dependent. We have included three different simulated datasets that satisfies all the assumptions underpinning a regression model. The aim is to use them as a standard for understanding the behaviours of other real-world datasets given a regression model. As we already know, the real-world datasets do not often comply with all the assumptions of a regression model. In that circumstance, how do the metrics namely MAE, MSE and RMSE behave on data comparatively to instances when all the assumptions are satisfied? The analysis that follows in the next section will help to answer these questions.

4. DATA ANALYSIS/RESULT PRESENTATION

A total of twelve different real-world and simulated datasets are going to be used in this study. The datasets are largely sourced from the internet and a description of each of them is contained in the section that follows:

4.1 Dataset Descriptions

Abalon Dataset: The dataset contains 4177 observations and 8 variables from zoology field. The variables consist of physical measurement used to determine age of abalone shell by cutting their cone and counting the rings, this is achieved using a microscope. The original data can be sourced from https://archive.ics.uci.edu/ml/datasets/abalone.

Bodyfat Dataset: The dataset is about estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. The source of the dataset is https://www.kaggle.com/ datasets/fedesoriano/body-fat-prediction-dataset.

Dataset_2196_cloud: These are data collected in a cloud-seeding experiment in Tasmania between mid-1964 and January 1971. Their analysis, using regression techniques and permutation tests, is discussed in (Miller et al., 1979). The dataset is contained on https://www.kaggle.com/code/milachadayammuri/clouddataviz/data.

Diamond Dataset: This dataset contains 53940 observation of 7 variables. It is about the analysis of diamonds by their cut, color, clarity, price, and other attributes. The source is https://www.kaggle.com/datasets/shivam2503/diamonds.

Fish Dataset: This dataset is a record of 7 common different fish species in fish market sales. It is used to perform a predictive model using machine friendly data and estimate of the weight of fish can be predicted. The source is https://www.kaggle.com/datasets/aungpyaeap/fish-market.

Garment worker productivity: The dataset consists of important attributes of the garment manufacturing process and the productivity of the employees which had been collected manually. It has 1197 observations and 10 attributes. The source is https://www.kaggle.com/datasets/ishadss/productivity-prediction-of-garment-employees.

Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 10, Issue 1, pp: (20-24), Month: January - April 2023, Available at: www.noveltyjournals.com

Wine Quality Dataset: The wine quality dataset concerns red and white variants of the Portuguese "Vinho Verde" wine (Cortez et al., 2009). As a result of privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available. The data can be downloaded from https://www.kaggle.com/datasets/rajyellow46/wine-quality?select=winequalityN.csv.

Salary Dataset: The dataset consists of one input variable and an outcome variable; hence it is suitable for carrying out simple regression modelling. There is no other documentation about the data and it can be downloaded from https://www.kaggle.com/datasets/karthickveerakumar/salary-data-simple-linear-regression.

Hungary Chickenpox: The dataset consists of a weekly chickenpox (childhood disease) cases from Hungary. It is made of county-level adjacency matrix and time series of the county-level reported cases between 2005 and 2015. There are two specific related tasks: County level case count prediction and nation level case count prediction. It can be downloaded from https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases#.

Simulated Dataset 1, 2 & 3: The datasets were simulated from a multivariate distribution but their variables were varied in all the cases. The R codes used to generate the datasets are contained in the Appendix.

4.2 Dataset Analysis/Result

The result of dataset analysis, using the R statistical software (R Core Team, 2018) is presented in Table 4.1.

Table 4.1: The outcome of data analysis showing the performances of five different performance metrices for regression, given the regression model of (1.1).

S/No.	Dataset	MAE	MSE	RMSE	R^2	$Adjusted\ R^2$
1	Abalone	1.53	4.71	2.17	0.495	0.492
2	Bodyfat	0.65	0.89	0.94	0.99	0.99
3	Dataset_2196_cloud	0.51	0.61	0.78	0.69	0.66
4	Diamond	408.17	268295	517.97	0.83	0.83
5	Fish Dataset	97.33	16197.03	127.27	0.86	0.84
6	Garment worker productivity	0.10	0.02	0.13	0.28	0.26
7	Wine quality	0.49	0.40	0.63	0.34	0.33
8	Simulated Dataset1 (dim 250 × 25)	2.31e-14	9.22e-28	3.04e-14	1	1
9	Simulated Dataset2 (dim 100 × 10)	1.44e-13	2.23e-26	1.49e-13	1	1
10	Salary Dataset	6094.26	47956403	6925.056	0.97	0.96
11	Simulated Dataset3 (dim 60 × 4)	1.12e-14	1.56e-28	1.25e-14	1	1
12	Hungarian_chickenpox	32.63	1991.62	44.63	0.65	0.60

Based on Table 4.1, MAE appears to have the smallest error relative to MSE and RMSE. It is followed by RMSE, whereas the MSE leaves behind errors that are larger comparatively. With respect to the garment workers productivity, MSE scored the least error among MAE and RMSE. Although the difference is clearly marginal, it goes further to suggest that the three metrices may be data dependent.

The R^2 and $Adjusted\ R^2$ have not shown any clear-cut difference because their individual values on each dataset are nearly identical. What may be seen as a surprise concerns the garment worker productivity dataset. Here, despite the fact that value for MAE, MSE and RMSE are all very low, yet the corresponding R^2 and $Adjusted\ R^2$ are equally very low. Ordinarily, one would have expected a high value for them. Similarly is the case of wine quality dataset and salary dataset. In the case of salary dataset, we observed very high values for MAE, MSE and RMSE, whereas their corresponding values for both R^2 and $Adjusted\ R^2$ are very high. Based on R^2 and $Adjusted\ R^2$ alone here, one would say that a corresponding model is rather optimum, but on giving consideration to other metrics, this point of view is strongly refuted.

In fact, considering the random behaviours of the performance metrics on data, it can be argued that the outcome of their performances is data dependent. Now, on the equality of the performances of MAE, MSE and RMSE on data, a non-parametric Kruskal-Wallis rank sum test (McKight & Najab, 2010) in R shows that at a p-value of 0.9298, a null hypothesis

of equality of the performances of the three metrics on data cannot be rejected. This, however, contradicts on-going argument that the metrices behave differently on data. In fact, based on the foregoing, their performances on data are statistically the same. Be that as it may, it is still reasonable to advise that the RMSE may be preferred because it is interpretable (Bobbitt, 2021).

On the other hand, a surprise recorded in this analysis concerns the simulated datasets. Here, it can be observed that all the performance metrices maintained optimal output. The errors associated with MAE, MSE and RMSE are all approximately zero, whereas their corresponding values for both R^2 and $Adjusted\ R^2$ are all one. Based on this outcome, it can be argued that in instances where all the assumptions underpinning a regression model are met (with the simulated datasets, all the assumptions of a regression model are met), any of the performance metrices can be optimal, otherwise, the RMSE may be used since it is interpretable.

5. SUMMARY/CONCLUSIONS

So far, we have comparatively assessed the performances of five performance metrics on data. We used nine real-world datasets and three simulated datasets. Analysis using the real-world datasets shows that the performances of MAE, MSE and RMSE on data are not data dependent. This means that irrespective of the dataset you are concerned with, any of the metrices can be optimal. We note that since the RMSE is interpretable, it may be preferred over others. We also discovered that it may not be a good practice to use any of MAE, MSE or RMSE alone without including either R^2 or $adjusted\ R^2$. For instance, with the garment worker productivity dataset, very small errors were recorded for MAE, MSE and RMSE and instead of observing a high value for the R^2 (by extension $adjusted\ R^2$), a small value was still recorded for them. A corresponding regression model in this instance may not be considered optimal. In such cases, a choice of another model may be made.

REFERENCES

[1] Beers, B. (2003). What is Regression? Definition, Calculation, and Example — investopedia.com https://www.investopedia.com/terms/r/regression.asp#:~:text=A%20regression%20is%20a%20statistical,more%20of%20the%20explanatory%20variables..

[2] Bobbitt, Z. (2021). How to Interpret Root Mean Square Error (RMSE) — statology.org. https://www.statology.org/how-to-interpret-rmse/

[3] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling Wine Preferences by Data Mining from Physicochemical Properties. Decision Support Systems, 47(4), 547–553.

[4] McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. The Corsini Encyclopedia of Psychology, 1.

[5] Miller, A. J., Shaw, D. E., Veitch, L. G., & Smith, E. J. (1979). Analyzing the Results of a Cloud-Seeding Experiment in Tasmania. Communications in Statistics-Theory and Methods, 8(10), 1017–1047.

[6] R Core Team. (2018). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

[7] Raj, R. (2022). Evaluation Metrics for Regression Models — enjoyalgorithms.com. https://www.enjoyalgorithms.com/blog/evaluation-metrics-regression-models