The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation

Davide Chicco[1], **Matthijs J. Warrens**[2], **and Giuseppe Jurman**[3]

[1]**University of Toronto**
[2]**University of Groningen**
[3]**Fondazione Bruno Kessler**

Corresponding author:
Davide Chicco[1]

Email address: davidechicco@davidechicco.it

ABSTRACT

Regression analysis makes up a large part of supervised machine learning, and consists ~~in~~ of the prediction of a continuous independent target from a set of other predictor variables. The difference between binary classification and regression is in the target range: in binary classification, the target can have only two values (usually encoded as 0 and 1), while in regression the target can have multiple values. Even if regression ~~is~~ analysis has been employed in a huge number of machine learning studies, no consensus has been reached on a single, unified, standard metric to assess the results of the regression itself. Many studies employ the mean square error (MSE) and its rooted variant (RMSE), or the mean absolute error (MAE) and its percentage variant (MAPE). Although useful, these rates share a common drawback: since their values can range between zero and +infinity, a single value of them does not say much about the performance of the regression with respect to the distribution of the ground truth elements. In this study, we focus on two rates that actually generate a high score only if the majority of the elements of a ground truth group has been correctly predicted: the coefficient of determination (R-squared) and the symmetric mean absolute percentage error (SMAPE). After showing their mathematical properties, we report a comparison between R^2 and SMAPE in several use cases and in ~~a~~ two real medical ~~scenario~~ scenarios. Our results demonstrate that the coefficient of determination (R-squared) is more informative and truthful than SMAPE, and does not have the interpretability limitations of MSE, RMSE, MAE, and MAPE. We therefore suggest the usage of R-squared as standard metric to evaluate regression analyses in any scientific domain.

1 INTRODUCTION

The role played by regression analysis in data science cannot be overemphasised: predicting a continuous target is a pervasive task not only in practical terms, but also at a conceptual level. Regression is deeply investigated even nowadays, to the point of still being worth of considerations in top journals (Jaqaman and Danuser, 2006; Altman and Krzywinski, 2015; Krzywinski and Altman, 2015), and ~~being~~ widespread used also in the current scientific war against COVID-19 (Chan et al., 2021; Raji and Lakshmi, 2020; Senapati et al., 2020; Gambhir et al., 2020). The theoretical basis of regression encompasses several aspects revealing hidden connections in the data and alternative perspectives even up to broadly speculative view: for instance, interpreting the whole statistical learning as a particular kind of regression (Berk, 2020), or framing deep neural networks as recursive generalised regressors (Wüthrich, 2020), or even provocatively pushing such considerations to the limit of setting the whole of statistics under the regression framework (Hannay, 2020). The relevancy of the topic ~~clearly reflects on the wide~~ is clearly reflected in the wide and heterogeneous literature covering the different aspects and insights of the regression analysis, from general overviews (Golberg and Cho, 2004; Freund et al., 2006; Montgomery et al., 2021),

to more technical studies (Sykes, 1993; Lane, 2002) or articles outlining practical applications (Draper and Smith, 1998; Rawlings et al., 2001; Chatterjee and Hadi, 2015), including handbooks (Chatterjee and Simonoff, 2013) or works covering specific key subtopics (Seber and Lee, 2012). However, the reference landscape is far wider: the aforementioned considerations stimulated a steady flow of studies investigating more philosophically oriented arguments (Allen, 2004; Berk, 2004), or deeper analysis of implications related to learning (Bartlett et al., 2020). Given the aforementioned overall considerations, it comes as no surprise that, similarly to what ~~happens~~ happened for binary classification, a plethora of performance metrics have been defined and are currently in use for evaluating the ~~goodness~~ quality of a regression model (Shcherbakov et al., 2013; Hyndman and Koehler, 2006; Botchkarev, 2018b,a, 2019). The parallel with classification goes even further: in the scientific community, a shared consensus on a preferential metric is indeed far from being reached, concurring to making comparison of methods and results a daunting task.

The present study provides a contribute towards the detection of critical factors in the choice of a suitable performance metric in regression analysis, through a comparative overview of two measures of current widespread use, namely the coefficient of determination and the symmetric mean absolute percentage error.

Indeed, despite the lack of a concerted standard, a set of well established and preferred metrics does exist and we believe that, as *primus inter pares*, the coefficient of determination R-squared deserves a major role. Introduced by Sewell Wright (1921) and generally indicated by R^2, ~~in~~ its original formulation ~~should quantify~~ quantifies how much the dependent variable is determined by the independent variables, in terms of proportion of variance. Again, given the age and diffusion of R^2, a wealth of studies about it has populated the scientific literature of the last century, from general references detailing definition and characteristics (Di Bucchianico, 2008; Barrett, 2000; Brown, 2009; Barrett, 1974), to more refined interpretative works (Saunders et al., 2012; Hahn, 1973; Nagelkerke, 1991; Ozer, 1985; Cornell and Berger, 1987; Quinino et al., 2013); efforts have been dedicated to the treatment of particular cases (Allen, 1997; Blomquist, 1980; Piepho, 2019; Srivastava et al., 1995; Dougherty et al., 2000; Cox and Wermuth, 1992; Zhang, 2017; Nakagawa et al., 2017; Menard, 2000) and to the proposal of *ad-hoc* variations (Young, 2000; Renaud and Victoria-Feser, 2010; Lee et al., 2012).

Parallel to the model explanation expressed as the variance, another widely adopted family of measures evaluate the ~~goodness~~ quality of fit in terms of distance of the regressor to the actual training points. The two basic members of such family are the mean average error (MAE) (Sammut and Webb, 2010a) and the mean squared error (MSE) (Sammut and Webb, 2010b), whose difference lies in the evaluating metric, respectively linear L_1 or quadratic L_2. Once more, the available references are numerous, related to both theoretical (David and Sukhatme, 1974; Rao, 1980; So et al., 2013) and applicative aspects (Allen, 1971; Farebrother, 1976; Gilroy et al., 1990; Imbens et al., 2005; Köksoy, 2006; Sarbishei and Radecka, 2011). As a natural derivation, the square root of mean square error (RMSE) has been widely adopted (Nevitt and Hancock, 2000; Hancock and Freeman, 2001; Applegate et al., 2003; Kelley and Lai, 2011) to standardize the units of measures of MSE. The different type of regularization imposed by the intrinsic metrics reflects on the relative effectiveness of the measure according to the data structure. In particular, as a rule of thumb, MSE is more sensitive to outliers than MAE; in addition to this general note, several further considerations helping ~~the~~ researchers in choosing the more suitable metric for evaluating a regression model given the available data and the target task can be drawn (Chai and Draxler, 2014; Willmott and Matsuura, 2005; Wang and Lu, 2018). Within the same family of measures, the mean absolute percentage error (MAPE) (de Myttenaere et al., 2016) focuses on the percentage error, being thus the elective metric when relative variations have a higher impact on the regression task rather than the absolute values. However, MAPE is heavily biased towards low forecasts, making it unsuitable for evaluating tasks where large errors are expected (Armstrong and Collopy, 1992; Ren and Glasure, 2009; De Myttenaere et al., 2015). Last but not least, the symmetric mean absolute percentage error (SMAPE) (Armstrong, 1985; Flores, 1986; Makridakis, 1993) is a recent metric originally proposed to solve some of the issues related to MAPE. Despite the yet not reached agreement on its optimal mathematical expression (Makridakis and Hibon, 2000; Hyndman and Koehler, 2006; Hyndman, 2014; Chen et al., 2017), SMAPE is progressively gaining momentum in the machine learning community due to its interesting properties (Maiseli, 2019; Kreinovich et al., 2014; Goodwin and Lawton, 1999),

An interesting discrimination among the aforementioned metrics can be formulated in terms of their output range. The coefficient of determination is upper bounded by the value 1, attained for perfect fit;

while R^2 is not lower bounded, the value 0 corresponds to (small perturbations of) the trivial fit provided by the horizontal line $y = K$ for K the mean of the target value of all the training ~~point~~ points. Since all negative values for R^2 indicate a worse fit than the average line, nothing is lost by considering the unit interval as the meaningful range for R^2. As a consequence, the coefficient of determination is invariant for linear transformations of the independent variables' distribution, and an output value close to one yields a good prediction regardless of the scale on which such variables are measured (Reeves, 2021). Similarly, also SMAPE values are bounded, with the lower bound 0% implying a perfect fit, and the upper bound 200% reached when all the predictions and the actual target values are of opposite sign. Conversely, MAE, MSE, RMSE and MAPE output spans the whole positive branch of the real line, with lower limit zero implying a perfect fit, and values progressively and infinitely growing for worse performing models. By definition, these values are heavily dependent on the describing variables' ranges, making them incomparable both mutually and within the same metric: a given output value for a metric has no interpretable relation with a similar value for a different measure, and even the same value for the same metric can reflect deeply different model performance for two distinct tasks (Reeves, 2021). Such property cannot be changed even if projecting the output into a bounded range through a suitable transformation (for example, arctangent or rational function). Given these interpretability issues, here we concentrate our comparative analysis on R^2 and SMAPE, both providing a high score only if the majority of the ground truth training points has been correctly predicted by the regressor. Showing the behaviour of these two metrics in several use cases and in ~~a~~ two biomedical ~~scenario~~ scenarios on ~~a dataset~~ two datasets made of ~~with 615~~ electronic health records ~~(75 hepatitis C patients and 540 healthy controls) described by 13 clinical factors~~, the coefficient of determination is demonstrated to be superior to SMAPE in terms of effectiveness and informativeness, thus being the recommended general performance measure to be used in evaluating regression analyses.

The manuscript organization proceeds as follows. After this Introduction, in the Methods section we introduce the cited metrics, with their mathematical definition and their main properties, and we provide a ~~deeper~~ more detailed description of ~~for~~ R^2 and SMAPE and their extreme values (section 2). In the following section Results and Discussion, we present the experimental part (section 3). First, we describe five synthetic use cases, then we introduce and detail the Lichtinghagen dataset and the Palechor dataset of electronic health records, together with the different applied regression models and the corresponding results. We complete that section ~~by~~ with a ~~the~~ discussion of the implication of all the obtained outcomes. In the Conclusions section, we draw some final considerations and future developments (section 4).

2 METHODS

In this section, we first introduce the mathematical background of the analyzed rates (subsection 2.1), then report some relevant information about the coefficient of determination and SMAPE (subsection 2.2).

2.1 Mathematical background

In the following formulas, X_i is the predicted i^{th} value, and the Y_i element is the actual i^{th} value. The regression method predicts the X_i element for the corresponding Y_i element of the ground truth dataset. Define two constants: the mean of the true values

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i \tag{1}$$

and the mean total sum of squares

$$\mathrm{MST} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y})^2 \tag{2}$$

Coefficient of determination (R^2 or R-squared)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \tag{3}$$

(worst value $= -\infty$; best value $= +1$)

The coefficient of determination (Wright, 1921) can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables.

Mean square error (MSE)

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \tag{4}$$

(best value $= 0$; worst value $= +\infty$)

MSE can be used if there are outliers that need to be detected. In fact, MSE is great for attributing larger weights to such points, thanks to the L_2 norm: clearly, if the model eventually outputs a single very bad prediction, the squaring part of the function magnifies the error.

Since $\text{R}^2 = 1 - \frac{\text{MSE}}{\text{MST}}$ and since MST is fixed for the data at hand, R^2 is monotonically related to MSE (a negative monotonic relationship), which implies that an ordering of regression models based on R^2 will be identical (although in reverse order) to an ordering of models based on MSE or RMSE.

Root mean square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{5}$$

(best value $= 0$; worst value $= +\infty$)

The two quantities MSE and RMSE are monotonically related (through the square root). An ordering of regression models based on MSE will be identical to an ordering of models based on RMSE.

Mean absolute error (MAE)

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m} \mid X_i - Y_i \mid \tag{6}$$

(best value $= 0$; worst value $= +\infty$)

MAE can be used if outliers represent corrupted parts of the data. In fact, MAE is not penalizing too much the training outliers (the L_1 norm somehow smooths out all the errors ~~on the outlier~~ of possible outliers), thus providing a generic and bounded performance measure for the model. On the other hand, if the test set also has many outliers ~~too~~, the model performance will be mediocre.

Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m} \left| \frac{Y_i - X_i}{Y_i} \right| \tag{7}$$

(best value $= 0$; worst value $= +\infty$)

MAPE is another performance metric for regression models, having a very intuitive interpretation in terms of relative error: due to its definition, its use is recommended in tasks where it is more important being sensitive to relative variations than to absolute variations (de Myttenaere et al., 2016). However, its has a number of drawbacks, too, the most critical ones being the restriction of its use to strictly positive data by definition and being biased towards low forecasts, which makes it unsuitable for predictive models where large errors are expected (Armstrong and Collopy, 1992).

Symmetric mean absolute percentage error (SMAPE)

$$\text{SMAPE} = \frac{100\%}{m} \sum_{i=1}^{m} \frac{| X_i - Y_i |}{(| X_i | + | Y_i |)/2} \tag{8}$$

(best value = 0; worst value = 2)

Initially defined by Armstrong (1985), and then refined in its current version by Flores (1986) and Makridakis (1993), SMAPE was proposed to amend the drawbacks of the MAPE metric. However, there is little consensus on a definitive formula for SMAPE, and different authors keep using slightly different versions (Hyndman, 2014). The original SMAPE formula defines the maximum value as 200%, which is computationally equivalent to 2. In this manuscript, we are going to use the first value for formal passages, and the second value for numeric calculations.

Informativeness The rates RMSE, MAE, MSE and SMAPE have value 0 if the linear regression model fits the data perfectly, and positive value if the fit is less than perfect. Furthermore, the coefficient of determination has value 1 if the linear regression model fits the data perfectly (that means if MSE = 0), value 0 if MSE = MST, and negative value if the mean squared error, MSE, is greater than mean total sum of squares, MST.

Even without digging into the mathematical properties of the aforementioned statistical rates, it is clear that it is difficult to interpret ~~that~~ sole values of MSE, RMSE, MAE, and MAPE, since they have $+\infty$ as upper bound. An MSE = 0.7, for example, does not say much about the overall quality of a regression model: the value could mean both an excellent regression model and a poor regression model. We cannot know it unless the maximum MSE value for the regression task is provided or unless the distribution of all the ground truth values is known. The same concept is valid for the other rates having $+\infty$ as upper bound, such as RMSE, MAE, and MAPE.

The only two regression scores that have strict real values are the non-negative R-squared and SMAPE. R-squared can have negative values, which mean that the regression performed poorly. R-squared can have value 0 when the regression model explains none of the variability of the response data around its mean (Minitab Blog Editor, 2013).

The positive values of the coefficient of determination range in the $[0,1]$ interval, with 1 meaning perfect prediction. On the other side, the values of SMAPE range in the $[0,2]$, with 0 meaning perfect prediction and 2 meaning worst prediction possible.

This is the main advantage of the coefficient of determination and SMAPE over RMSE, MSE, MAE, and MAPE: values like R^2 = 0.8 and SMAPE = 0.1, for example, clearly indicate a very good regression model performance, ~~despite~~ regardless of the ranges of the ground truth values and their distributions. A value of RMSE, MSE, MAE, or MAPE equal to 0.7, instead, fails to inform us about the quality of the regression performed.

This property of R-squared and SMAPE can be useful in particular when one needs to compare the predictive performance of a regression on two different datasets having different value scales. For example, suppose we have a mental health study describing a predictive model where the outcome is a depression scale ranging from 0 to 100, and another study using a different depression scale, ranging from 0 to 10 (Reeves, 2021). Using R-squared or SMAPE we could compare the predictive performance of the two studies without making additional transformations. The same comparison would be impossible with RMSE, MSE, MAE, or MAPE.

Given the ~~superiority~~ better robustness of R-squared and SMAPE over the other four rates, we focus the rest of this article on the comparison between ~~them~~ these two statistics.

2.2 R-squared and SMAPE

R-squared The coefficient of determination can take values in the range $(-\infty, 1]$ according to the mutual relation between the ground truth and the prediction model. Hereafter we report a brief overview of the principal cases.

$R^2 \geq 0$: With linear regression with no constraints, R^2 is non-negative and corresponds to the square of the multiple correlation coefficient.

$R^2 = 0$: The fitted line (or hyperplane) is horizontal. With two numerical variables this is the case if the variables are independent, that is, are uncorrelated. Since $R^2 = 1 - \frac{MSE}{MST}$, the relation $R^2 = 0$ is equivalent to MSE = MST, or, equivalently, to:

$$\sum_{i=1}^{m}(Y_i - \bar{Y})^2 = \sum_{i=1}^{m}(Y_i - X_i)^2 \tag{9}$$

Now, Equation 9 has the obvious solution $X_i = \bar{Y}$ for $1 \leq i \leq m$, but, being just one quadratic equation with m unknowns X_i, it has infinite solutions, where $X_i = \bar{Y} \pm \varepsilon_i$ for a small ε_i, as shown in the following example:

- $\{Y_i\, 1 \leq i \leq 10\} = \{90.317571, 40.336481, 5.619065, 44.529437, 71.192687, 32.036909, 6.977097, 66.425010, 95.971166, 5.756337\}$
- $\bar{Y} = 45.91618$
- $\{X_i\, 1 \leq i \leq 10\} = \{45.02545, 43.75556, 41.18064, 42.09511, 44.85773, 44.09390, 41.58419, 43.25487, 44.27568, 49.75250\}$
- MSE = MST = 1051.511
- $R^2 \approx 10^{-8}$.

$R^2 < 0$: This case is only possible with linear regression when either the intercept or the slope are constrained so that the "best-fit" line (given the constraint) fits worse than a horizontal line, for instance if the regression line (hyperplane) does not follow the data (CrossValidated, 2011b). With nonlinear regression, the R-squared can be negative whenever the best-fit model (given the chosen equation, and its constraints, if any) fits the data worse than a horizontal line. Finally, negative R^2 might also occur when omitting a constant from the equation, that is, forcing the regression line to go through the point (0,0).

A final note. The behavior of the coefficient of determination is rather independent from the linearity of the regression fitting model: R^2 can be very low even for completely linear model, and vice versa, a high R^2 can occur even when the model is noticeably non-linear. In particular, a good global R^2 can be split in several local models with low R^2 (CrossValidated, 2011a).

SMAPE By definition, SMAPE values range between 0% and 200%, where the following holds in the two extreme cases:

SMAPE = 0: The best case occurs when SMAPE vanishes, that is when

$$\frac{100\%}{m}\sum_{i=1}^{m}\frac{|\,X_i - Y_i\,|}{(|\,X_i\,| + |\,Y_i\,|)/2} = 0$$

equivalent to

$$\sum_{i=1}^{m}\frac{|\,X_i - Y_i\,|}{(|\,X_i\,| + |\,Y_i\,|)/2} = 0$$

and, since the m components are all positive, equivalent to

$$\frac{|\,X_i - Y_i\,|}{|\,X_i\,| + |\,Y_i\,|} = 0 \;\forall\; 1 \leq i \leq m$$

and thus $X_i = Y_i$, that is, perfect regression.

SMAPE = 2: The worst case SMAPE = 200% occurs instead when

$$\frac{100\%}{m}\sum_{i=1}^{m}\frac{|\,X_i - Y_i\,|}{(|\,X_i\,| + |\,Y_i\,|)/2} = 2$$

equivalent to

$$\sum_{i=1}^{m} \frac{|\, X_i - Y_i \,|}{|\, X_i \,| + |\, Y_i \,|} = m$$

By the triangle inequality $|\, a+c \,| \leq |\, a \,| + |\, c \,|$ computed for $b = -c$, we have that $|\, a-b \,| \leq |\, a \,| + |\, b \,|$, and thus $\frac{|a-b|}{|a|+|b|} \leq 1$. This yields that SMAPE = 2 if $\frac{|X_i - Y_i|}{|X_i|+|Y_i|} = 1$ for all $i = 1, \ldots, m$. Thus we reduced to compute when $\xi(a,b) = \frac{|a-b|}{|a|+|b|} = 1$: we analyse now all possible cases, also considering the symmetry of the relation with respect to a and b, $\xi(a,b) = \xi(b,a)$.

If $a = 0$, $\xi(0,b) = \frac{|0-b|}{|0|+|b|} = 1$ if $b \neq 0$.

Now suppose that $a, b > 0$: $\xi(a,a) = 0$, so we can suppose $a > b$, thus $a = b + \varepsilon$, with $a, b, \varepsilon > 0$. Then $\xi(a,b) = \xi(b+\varepsilon, \varepsilon) = \frac{\varepsilon}{2b+\varepsilon} < 1$. Same happens when $a, b < 0$: thus, if ground truth points and the prediction points have the same sign, SMAPE will never reach its maximum value.

Finally, suppose that a and b have opposite sign, for instance $a > 0$ and $b < 0$. Then $b = -c$, for $c > 0$ and thus $\xi(a,b) = \xi(a,-c) = \frac{|a+c|}{|a|+|c|} = \frac{a+c}{a+c} = 1$.

Summarising, SMAPE reaches its worst value 200% if

- $X_i = 0$ and $Y_i \neq 0$ for all $i = 1, \ldots, m$
- $X_i \neq 0$ and $Y_i = 0$ for all $i = 1, \ldots, m$
- $X_i \cdot Y_i < 0$ for all $i = 1, \ldots, m$, that is, ground truth and prediction always have opposite sign, regardless of their values.

For instance, if the ground truth points are `(1, -2, 3, -4, 5, -6, 7, -8, 9, -10)`, any prediction vector with all opposite signs (for example, `(-307.18, 636.16, -469.99, 671.53, -180.55, 838.23, -979.18 , 455.16, -8.32, 366.80)`) will result in a SMAPE metric reaching 200%.

Explained the extreme cases of R-squared and SMAPE, in the next section we illustrate some significant, informative use cases where these two rates generate discordant outcomes.

3 RESULTS AND DISCUSSION

In this section, we first report some particular use cases where we compare the results of R-squared and SMAPE (subsection 3.1), and then we describe a real biomedical scenario where the analyzed regression rates generate different rankings for the methods involved (subsection 3.2).

As mentioned earlier, we exclude MAE, MSE, RMSE, and MAPE from the selection of the best performing regression rate. These statistics range in the $[0, +\infty)$ interval, with 0 meaning perfect regression, and their values alone therefore fail to communicate the quality of the regression performance, both on good cases and in bad cases. We know for example that a negative coefficient of determination and a SMAPE equal to 1.9 clearly correspond to a regression which performed poorly, but we do not have a specific value for MAE, MSE, RMSE, and MAPE that indicates this outcome. Moreover, as mentioned earlier, each value of MAE, MSE, RMSE, and MAPE communicates the quality of the regression only relatively to other regression performances, and not in an absolute manner, like R-squared and SMAPE do. For these reasons, we focus on the coefficient of determination and SMAPE for the rest of our study.

3.1 Use cases

We list hereafter a number of example use cases where the coefficient of determination and SMAPE produce divergent outcomes, showing that R^2 is more robust and reliable than SMAPE, especially on bad regressions. To simplify comparison between the two measures, define the complementary normalized SMAPE as:

$$\text{cnSMAPE} = 1 - \frac{\text{SMAPE}}{200\%} \tag{10}$$

(worst value = 0; best value = 1)

UC1 Use case Consider the ground truth set $\text{REAL} = \{r_i = (i,i) \in \mathbb{R}^2,\ i \in \mathbb{N}, 1 \leq i \leq 100\}$ collecting 100 points with positive integer coordinates on the straight line $y = x$. Define then the set $\text{PRED}_j = \{p_i\}$ as

$$p_i = \begin{cases} r_i & \text{if } i \not\equiv 1 \pmod 5 \\ r_{5k+1} & \text{for } k \geq j \\ 0 & \text{for } i = 5k+1\ , 0 \leq k < j \end{cases} \tag{11}$$

so that REAL and PRED_j coincides apart from the first j points $1, 6, 11, \ldots$ congruent to 1 modulo 5 that are set to 0. Then, for each $5 \leq j \leq 20$, compute R^2 and cnSMAPE (Table 1).

Table 1. UC1 Use case. Values generated through Equation 11. R^2: coefficient of determination (Equation 3). cnSMAPE: complementary normalized SMAPE (Equation 10).

Both measures decrease with the increasing number of non-matching points $p_{5k+1} = 0$, but cnSMAPE decreases linearly, while R^2 goes down much faster, better showing the growing unreliability of the predicted regression. At the end of the process, $j = 20$ points out of 100 are wrong, but still cnSMAPE is as high as 0.80, while R^2 is 0.236, correctly declaring PRED_{20} a very weak prediction set.

UC2 Use case In a second example, consider again the same REAL dataset and define the three predicting sets

$$\text{PRED}_{\text{start}} = \{p_i^s \colon 1 \leq i \leq 100\}$$
$$p_i^s = \begin{cases} r_i & \text{for } i \geq 10 \\ 0 & \text{for } i < 10 \end{cases}$$
$$\text{PRED}_{\text{middle}} = \{p_i^m \colon 1 \leq i \leq 100\}$$
$$p_i^m = \begin{cases} r_i & \text{for } i \leq 50 \text{ and } i \geq 61 \\ 0 & \text{for } 51 \leq i \leq 60 \end{cases}$$
$$\text{PRED}_{\text{end}} = \{p_i^e \colon 1 \leq i \leq 100\}$$
$$p_i^e = \begin{cases} r_i & \text{for } i \leq 90 \\ 0 & \text{for } i \geq 91 \end{cases}$$

In all the three cases *start*, *middle*, *end* the predicting set coincides with REAL up to 10 points that are set to zero, at the beginning, in the middle and at the end of the prediction, respectively. Interestingly, cnSMAPE is 0.9 in all the three cases, showing that SMAPE is sensible only to the number of non-matching points, and not to the magnitude of the predicting error. R^2 instead correctly decreases when the zeroed sequence of points is further away in the prediction and thus farthest away from the actual values: R^2 is 0.995 for $\text{PRED}_{\text{start}}$, 0.6293 for $\text{PRED}_{\text{middle}}$ and –0.0955 for PRED_{end}.

UC3 Use case Consider now the as the ground truth the line $y = x$, and sample the set T including twenty positive integer points $T = \{t_i = (x_i, y_i^T) = (i,i)\ 1 \leq i \leq 20\}$ on the line. Define $\text{REAL} = \{r_i = (x_i, y_i^R) = (i, i + N(i))\ 1 \leq i \leq 20\}$ as the same points of T with a small amount of noise $N(i)$ on the y axes, so that r_i are close but not lying on the $y = x$ straight line. Consider now two predicting regression models:

- The set $\text{PRED}_c = T$ representing the correct model;
- The set PRED_w representing the (wrong) model with points defined as $p_i^w = f(x_i)$, for f the 10-th degree polynomial exactly passing through the points r_i for $1 \leq i \leq 10$.

Clearly, p_i^w coincides with r_i for $1 \leq i \leq 10$, but $\|p_i^w - r_i\|$ becomes very large for $i \geq 11$. On the other hand $t_i \neq r_i$ for all i's, but $\|t_i - r_i\|$ is always very small. Compute now the two measures R^2 and cnSMAPE

Table 2. UC3 Use case. We define N, correct model, and wrong model in the UC3 Use case paragraph. R^2: coefficient of determination (Equation 3). cnSMAPE: complementary normalized SMAPE (Equation 10).

on the first N points $i = 1,\ldots,N$ for $2 \leq N \leq 20$ of the two different regression models c and w with respect to the ground truth set REAL (Table 2).

For the correct regression model, both measures are correctly showing good results. For the wrong model, both measures are optimal for the first 10 points, where the prediction exactly matches the actual values; after that, R^2 rapidly decreases supporting the inconsistency of the model, while cnSMAPE is not affected that much, arriving for $N = 20$ to a value $1/2$ as a minimum, even if the model is clearly very bad in prediction.

UC4 Use case Consider the following example: the seven actual values are `(1, 1, 1, 1, 1, 2, 3)`, and the predicted values are `(1, 1, 1, 1, 1, 1, 1)`. From the predicted values, it is clear that the regression method worked very poorly: it predicted `1` for all the seven values.

If we compute the coefficient of determination and SMAPE here, we obtain R-squared = –0.346 and SMAPE = 0.238. The coefficient of determination illustrates that something is completely off, by having a negative value. On the contrary, SMAPE has a very good score, that corresponds to 88.1% correctness in the cnSMAPE scale.

In this use case, if a inexperienced practitioner decided to check only the value of SMAPE to evaluate her/his regression, she/he would be misled and would wrongly believe that the regression went 88.1% correct. If, instead, the practitioner decided to verify the value of R-squared, she/he would be alerted about the poor quality of the regression. As we saw earlier, the regression method predicted `1` for all the seven ground truth elements, so it clearly performed poorly.

UC5 Use case Let us consider now a vector of 5 integer elements having values `(1, 2, 3, 4, 5)`, and a regression prediction made by the variables `(a, b, c, d, e)`. Each of these variables can assume all the integer values between 1 and 5, included. We compute the coefficient of determination and cnSMAPE for each of the predictions with respect to the actual values. To compare the values of the coefficient of determination and cnSMAPE in the same range, we consider only the cases when R-squared is greater or equal to zero, and we call it non-negative R-squared. We reported the results in Figure 1.

Figure1_examples_Rsquared_cnSMAPE_five_elements.png

Figure 1. UC5 Use case: R-squared versus cnSMAPE. Representation plot of the values of cnSMAPE (Equation 10) on the y axis and non-negative R-squared (Equation 3) on the x axis, obtained in the UC5 Use case. Blue line: regression line generated with the *loess* smooth method.

As clearly observable in the plot Figure 1, there are a number of points where cnSMAPE has a high value (between 0.6 and 1) but R-squared had value 0: in these cases, the coefficient of determination and cnSMAPE give discordant outcomes. One of these cases, for example, is the regression where the predicted values have values `(1, 2, 3, 5, 2)`, $R^2 = 0$, and cnSMAPE = 0.89.

In this example, cnSMAPE has a very high value, meaning that the prediction is 89% correct, while R^2 is equal to zero. The regression correctly predicts the first three points `(1, 2, 3)`, but fails to classify the forth element (`4` is wrongly predicted as `5`), and the fifth element (`5` is mistakenly labeled as `2`). The coefficient of determination assigns a bad outcome to this regression because it fails to correctly classify the only members of the `4` and `5` classes. Diversely, SMAPE assigns a good outcome to this prediction because the variance between the actual values and the predicted values is low, in proportion to the overall mean of the values.

Faced with this situation, we consider the outcome of the coefficient of determination more reliable and trustworthy: similarly to the Matthews correlation coefficient (MCC) (Matthews, 1975) in binary classification (Chicco and Jurman, 2020; Chicco et al., 2021; Tötsch and Hoffmann, 2021; Chicco et al., 2021), R-squared generates a high score only if the regression is able to correctly classify most of the

elements of each class. In this example, the regression fails to classify all the elements of the 4 class and of the 5 class, so we believe a good metric would communicate this key-message.

3.2 Medical scenarios

To further investigate the behavior of R-squared, MAE, MAPE, MSE, RMSE, and SMAPE, we employed these rates to a regression analysis applied to ~~a~~ two real biomedical ~~application~~ applications.

Hepatitis dataset We trained and applied several machine learning regression methods on the Lichtinghagen dataset (Lichtinghagen et al., 2013; Hoffmann et al., 2018), which consists of electronic health records of 615 individuals including healthy controls and patients diagnosed with cirrhosis, fibrosis, and hepatitis. This dataset has 13 features, including a numerical variable stating the diagnosis of the patient, and is publicly available in the University of California Irvine Machine Learning Repository (2020). There are 540 healthy controls (87.8%) and 75 patients diagnosed with hepatitis C (12.2%). Among the 75 patients diagnosed with hepatitis C, there are: 24 with only hepatitis C (3.9%); 21 with hepatitis C and liver fibrosis (3.41%); and 30 with hepatitis C, liver fibrosis, and cirrhosis (4.88%)

Obesity dataset To further verify the effect of the regression rates, we applied the data mining methods to another medical dataset made of electronic health records of young patients with obesity (Palechor and De-La-Hoz-Manotas, 2019; De-La-Hoz-Correa et al., 2019). This dataset is publicly available in the University of California Irvine Machine Learning Repository (2019) too, and contains data of 2,111 individuals, with 17 variables for each of them. A variable called *NObeyesdad* indicates the obesity level of each subject, and can be employed as a regression target. In this dataset, there are 272 children with insufficient weight (12.88%), 287 children with normal weight (13.6%), 351 children with obesity type I (16.63%), 297 children with obesity type II (14.07%), 324 children with obesity type III (15.35%), 290 children with overweight level I (13.74%), and 290 children with overweight level II (13.74%). The original curators synthetically generated part of this dataset (Palechor and De-La-Hoz-Manotas, 2019; De-La-Hoz-Correa et al., 2019).

Methods For the regression analysis, we employed the same machine learning methods ~~the original~~ two of us authors used in ~~their~~ a previous analysis (Chicco and Jurman, 2021): Linear Regression (Montgomery et al., 2021), Decision Trees (Rokach and Maimon, 2005), and Random Forests (Breiman, 2001), all implemented and executed in the R programming language (Ihaka and Gentleman, 1996). For each method execution, we first shuffled the patients data, and then we randomly selected 80% of the data elements for the training set and used the remaining 20% for the test set. We trained each method model on the training set, applied the trained model to the test set, and saved the regression results measured through R-squared, MAE, MAPE, MSE, RMSE, and SMAPE. For the hepatitis dataset, we imputed the missing data with the Predictive Mean Matching (PMM) approach through the Multiple Imputation by Chained Equations (MICE) ~~software package~~ method (Buuren and Groothuis-Oudshoorn, 2010). We ran 100 executions and reported the results means and the rankings based on the different rates in Table 3 (hepatitis dataset) and in Table 4 (obesity dataset) .

Hepatitis dataset results: different rate, different ranking We measured the results obtained by these regression models on the Lichtinghagen hepatitis dataset with all the rates analyzed in our study: R^2, MAE, MAPE, RMSE, MSE, and SMAPE (lower part of Table 3).

These rates generate 3 different rankings. R^2, MSE, and RMSE share the same ranking (Random Forests, Linear Regression, and Decision Tree). SMAPE and MAPE share the same ranking (Decision Tree, Random Forests, and Linear Regression). MAE has its own ranking (Random Forests, Decision Tree, and Linear Regression).

It is also interesting to notice that these six rates select different methods as top performing method. R^2, MAE, MSE, and RMSE indicate Random Forests as top performing regression model, while SMAPE and MAPE select Decision Tree for the first position in their rankings. The position of Linear Regression changes, too: on the second rank for R^2, MSE, and RMSE, while on the last rank for MAE, SMAPE, and MAPE.

By comparing all these different standings, a machine learning practitioner could wonder what is the most suitable rate to choose, to understand how the regression experiments actually went and which method outperformed the others. As explained earlier, we suggest the readers to focus on the ranking generated by the coefficient of determination, because it is the only metric that considers the distribution

Table 3. Regression results on the prediction of hepatitis, cirrhosis, and fibrosis from electronic health records, and corresponding rankings based on rates. We performed the analysis on the Lichtinghagen dataset (Lichtinghagen et al., 2013; Hoffmann et al., 2018) with the methods employed by Chicco and Jurman (2021). We report here the average values achieved by each method in 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. R^2: worst value $-\infty$ and best value +1. SMAPE: worst value 2 and best value 0. MAE, MAPE, MSE, and RMSE: worst value $+\infty$ and best value 0. We reported the complete regression results including the standard deviations in Table S1. R^2 formula: Equation 3. MAE formula: Equation 6. MAPE formula: Equation 7. MSE formula: Equation 4. RMSE formula: Equation 5. SMAPE formula: Equation 8.

Table 4. Regression results on the prediction of obesity level from electronic health records, including standard deviations. Mean values and standard deviations out of 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. We performed the analysis on the Palechor dataset (Palechor and De-La-Hoz-Manotas, 2019; De-La-Hoz-Correa et al., 2019) with the methods Linear Regression, Decision Tree, and Random Forests. We report here the average values achieved by each method in 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. R^2: worst value $-\infty$ and best value +1. SMAPE: worst value 2 and best value 0. MAE, MAPE, MSE, and RMSE: worst value $+\infty$ and best value 0. We reported the complete regression results including the standard deviations in Table S2. R^2 formula: Equation 3. MAE formula: Equation 6. MAPE formula: Equation 7. MSE formula: Equation 4. RMSE formula: Equation 5. SMAPE formula: Equation 8.

of all the ground truth values, and generates a high score only if the regression correctly predict most of the values of each ground truth category. Additionally, the fact that the ranking indicated by R-squared (Random Forests, Linear Regression, and Decision Tree) was the same standing generated by 3 rates out of 6 suggests that it is the most informative one (Table 3).

Hepatitis dataset results: R^2 provides the most informative outcome Another interesting aspect of these results ~~regard~~ on the hepatitis dataset regards the comparison between coefficient of determination and SMAPE (Table 3). We do not compare the standing of R-squared with MAE, MSE, RMSE, and MAPE because these four rates can have infinite positive values and, as mentioned earlier, this aspect makes it impossible to detect the quality of a regression from a single score of these rates.

R-squared ~~generates~~ indicates a very good result for Random Forests ($R^2 = 0.756$), and good results for Linear Regression ($R^2 = 0.535$) and Decision Tree ($R^2 = 0.423$). On the contrary, SMAPE generates an excellent result for Decision Tree (SMAPE = 0.073), meaning almost perfect prediction, and poor results for Random Forests (SMAPE = 1.808) and Linear Regression (SMAPE = 1.840), very close to the upper bound (SMAPE = 2) representing the worst possible regression.

These values mean that the coefficient of determination and SMAPE generate discordant outcomes for these two methods: for R-squared, Random Forests made a very good regression and Decision Tree made a good one; for SMAPE, instead, Random Forests made a catastrophic regression and Decision Tree made an almost perfect one. At this point, a practitioner could wonder which algorithm between Random Forests and Decision Trees made the better regression. Checking the standings of the other rates, we clearly see that Random Forests resulted being the top model for 4 rates out of 6, while Decision Tree resulted being the worst model for 3 rates out of 6. This information confirms that the ranking of R-squared is more reliable than the one of SMAPE (Table 3).

Obesity dataset results: agreement between rankings, except for SMAPE Differently from the rankings generated on the hepatitis dataset, the rankings produced on the obesity dataset are more concordant (Table 4). Actually, the ranking of the coefficient of determination, MSE, RMSE, MAE, and MAPE are identical: Random Forests on the first position, Decision Tree on the second position, and

Linear Regression on the third and last position. All the rates' rankings indicate Random Forests as the top performing method.

The only significant difference can be found in the SMAPE standing: differently from the other rankings that all put Decision Tree as second best regressor and Linear Regression as worst regressor, the SMAPE standing indicates Linear Regression as runner-up and Decision Tree on the last position. SMAPE, in fact, swaps the positions of these two methods, compared to R-squared and the other rates: SMAPE says Linear Regression outperformed Decision Tree, while the other rates say that Decision Tree outperformed Linear Regression.

Since five out of six rankings confirm that Decision Tree generated better results than Linear Regression, and only one of six say vice versa, we believe that is clear that the ranking indicated by the coefficient of determination is more informative and trustworthy than the ranking generated by SMAPE.

4 CONCLUSIONS

Even if regression analysis makes a big chunk of the whole machine learning and computational statistics domains, no consensus has been reached on a unified prefered rate to evaluate regression analyses yet. In this study, we compared several statistical rates commonly employed in the scientific literature for regression task evaluation, and described the advantages of R-squared over SMAPE, MAPE, MAE, MSE, and RMSE.

~~Even if MAPE, MAE, MSE, and RMSE are still employed often in several studies~~ Despite the fact that MAPE, MAE, MSE, and RMSE are commonly used in machine learning studies, we showed that it is impossible to detect the quality of the performance of a regression method by just looking at their singular values. An MAPE of 0.7 alone, for example, fails to communicate if the regression algorithm performed mainly correctly or poorly. This ~~big~~ flaw left room only for R^2 and SMAPE. The first one has negative values if the regression performed poorly, and values between 0 and 1 (included) if the regression was good. A positive value of R-squared can be considered similar to percentage of correctness obtained by the regression. SMAPE, instead, has the value 0 as best value for perfect regressions and has the value 2 as worst value for disastrous ones.

In our study, we showed with several use cases and examples that R^2 is more truthful and informative ~~of~~ than SMAPE: R-squared, in fact, generates a high score only if the regression correctly predicted most of the ground truth elements for each ground truth group, considering their distribution. SMAPE, instead, focuses on the relative distance between each predicted value and its corresponding ground truth element, without considering their distribution. In the present study SMAPE turned out to perform bad in identifying bad regression models.

A limitation of R^2 arises in the negative space. When R-squared has negative values, it ~~says~~ indicates that the model performed poorly but it is impossible to know how bad a model performed. For example, an R-squared equal to –0.5 alone does not say much about the quality of the model, because the lower bound is $-\infty$. Differently from SMAPE that has values between 0 and 2, the minus sign of the coefficient of determination would however clearly inform the practitioner about the poor performance of the regression.

Although regression analysis can be applied to an infinite number of different datasets, with infinite values, we had to limit the present to a selection of cases, for feasibility purposes. The selection of use cases presented here are to some extent limited, since one could consider infinite many other use cases that we could not analyze here. Nevertheless, we did not find any use cases in which SMAPE turned out to be more informative than R-squared. Based on the results of this study and our own experience, R-squared seems to be the most informative rate in many cases, if compared to SMAPE, MAPE, MAE, MSE, and RMSE. We therefore suggest the employment of R-squared as the standard statistical measure to evaluate regression analyses, in any scientific area.

In the future, we plan to compare R^2 with other regression rates such as Huber metric H_δ (Huber, 1992), LogCosh loss (Wang et al., 2020), and Quantile Q_γ (Yue and Rue, 2011).

LIST OF ABBREVIATIONS

COVID-19: coronavirus disease 2019. DT: Decision Trees. LR: Linear Regression. MAE: mean absolute error. MAPE: Mean absolute percentage error. MSE: mean square error. R^2: R-squared, coefficient of determination. RF: Random Forests. RMSE: root mean square error. SMAPE: symmetric mean absolute percentage error.

SOFTWARE AVAILABILITY

Our software code is publicly available under GNU General Public License v3.0 at: `https://github.com/davidechicco/R-squared_versus_other_regression_rates`

REFERENCES

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475.

Allen, M. P. (1997). The coefficient of determination in multiple regression. *Understanding Regression Analysis*, pages 91–95.

Allen, M. P. (2004). *Understanding Regression Analysis*. Springer Science & Business Media, Berlin, Germany.

Altman, N. and Krzywinski, M. (2015). Simple linear regression. *Nature Methods*, 12(11):999–1000.

Applegate, R. A., Ballentine, C., Gross, H., Sarver, E. J., and Sarver, C. A. (2003). Visual acuity as a function of Zernike mode and level of root mean square error. *Optometry and Vision Science*, 80(2):97–105.

Armstrong, J. S. (1985). *Long-Range Forecasting: from Crystal Ball to Computer*. John Wiley & Sons, Hoboken, New Jersey, USA.

Armstrong, J. S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, 08:69–80.

Barrett, G. B. (2000). The coefficient of determination: understanding r^2 and R^2. *The Mathematics Teacher*, 93(3):230–234.

Barrett, J. P. (1974). The coefficient of determination – some limitations. *The American Statistician*, 28(1):19–20.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences of the USA*, 117(48):30063–30070.

Berk, R. A. (2004). *Regression Analysis: a Constructive Critique*, volume 11. Sage, Thousand Oaks, California, USA.

Berk, R. A. (2020). Statistical learning as a regression problem. In *Statistical Learning from a Regression Perspective*, pages 1–72. Springer International Publishing.

Blomquist, N. S. (1980). A note on the use of the coefficient of determination. *Scandinavian Journal of Economics*, 82(3):409–412.

Botchkarev, A. (2018a). Evaluating performance of regression machine learning models using multiple error metrics in Azure machine learning studio. *SSRN Electronic Journal*, 12 May 2018:3177507.

Botchkarev, A. (2018b). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. *arXiv*, 1809.03006:1–37.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:045–076.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, J. D. (2009). The coefficient of determination. `https://hosted.jalt.org/test/bro_16.htm`. URL visited on 22nd January 2021.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, pages 1–68.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chan, S., Chu, J., Zhang, Y., and Nadarajah, S. (2021). Count regression models for COVID-19. *Physica A: Statistical Mechanics and its Applications*, 563:125460.

Chatterjee, S. and Hadi, A. S. (2015). *Regression Analysis by Example*. John Wiley & Sons, Hoboken, New Jersey, USA.

Chatterjee, S. and Simonoff, J. S. (2013). *Handbook of Regression Analysis*, volume 5. John Wiley & Sons, Hoboken, New Jersey, USA.

Chen, C., Twycross, J., and Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, 12(3):e0174202.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13.

Chicco, D. and Jurman, G. (2021). An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. *IEEE Access*, 9:24485–24498.

Chicco, D., Starovoitov, V., and Jurman, G. (2021). The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access*, 9:47112–47124.

Chicco, D., Tötsch, N., and Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):1–22.

Cornell, J. A. and Berger, R. D. (1987). Factors that influence the value of the coefficient of determination in simple linear and nonlinear regression models. *Phytopathology*, 77(1):63–70.

Cox, D. R. and Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *The American Statistician*, 46(1):1–4.

CrossValidated (2011a). Is R^2 useful or dangerous? `https://stats.stackexchange.com/questions/13314/is-r2-useful-or-dangerous/13317#13317` URL visited on 23rd February 2021.

CrossValidated (2011b). When is R squared negative? `https://stats.stackexchange.com/questions/12900/when-is-r-squared-negative` URL visited on 19th February 2021.

David, I. P. and Sukhatme, B. V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association*, 69(346):464–466.

De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C., and Adriana, S. H. B. (2019). Obesity level estimation software based on decision trees. *Journal of Computer Science*, 15(1):67–77.

De Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2015). Using the mean absolute percentage error for regression models. In *Proceedings of ESANN 2015 – the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, page 113. Presses Universitaires de Louvain.

de Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48.

Di Bucchianico, A. (2008). Coefficient of determination (R^2). *Encyclopedia of Statistics in Quality and Reliability*, 1.

Dougherty, E. R., Kim, S., and Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80(10):2219–2235.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, volume 326. John Wiley & Sons, Hoboken, New Jersey, USA.

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–250.

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98.

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression Analysis*. Elsevier, Amsterdam, Netherlands.

Gambhir, E., Jain, R., Gupta, A., and Tomer, U. (2020). Regression analysis of COVID-19 using machine learning algorithms. In *Proceedings of ICOSEC 2020 – the 2nd International Conference on Smart Electronics and Communication*, pages 65–71. IEEE.

Gilroy, E. J., Hirsch, R. M., and Cohn, T. A. (1990). Mean square error of regression-based constituent transport estimates. *Water Resources Research*, 26(9):2069–2077.

Golberg, M. A. and Cho, H. A. (2004). *Introduction to Regression Analysis*. WIT Press, Ashurst, New Forest, England, United Kingdom.

Goodwin, P. and Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, 15(4):405–408.

Hahn, G. J. (1973). The coefficient of determination exposed. *Chemical Technology*, 3(10):609–612.

Hancock, G. R. and Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, 61(5):741–758.

Hannay, K. (2020). Everything is a regression: in search of unifying paradigms in statistics. `https://towardsdatascience.com/everything-is-just-a-regression-5a3bf22c459c` URL visited on 15th March 2021. Towards Data Science.

Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., and Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways – a case study. *Journal of Laboratory and Precision Medicine*, 3:58.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in Statistics*, pages 492–518. Springer.

Hyndman, R. J. (2014). Errors on percentage errors. `https://robjhyndman.com/hyndsight/smape/`. URL visited on 26th February 2021.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Imbens, G. W., Newey, W. K., and Ridder, G. (2005). Mean-square-error calculations for average treatment effects. Technical report, Harvard University.

Jaqaman, K. and Danuser, G. (2006). Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11):813–819.

Kelley, K. and Lai, K. (2011). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research*, 46(1):1–32.

Köksoy, O. (2006). Multiresponse robust design: mean square error (MSE) criterion. *Applied Mathematics and Computation*, 175(2):1716–1729.

Kreinovich, V., Nguyen, H. T., and Ouncharoen, R. (2014). How to estimate forecasting quality: a system-motivated derivation of symmetric mean absolute percentage error (SMAPE) and other similar characteristics. Technical Report UTEP-CS-14-53, University of Texas at El Paso.

Krzywinski, M. and Altman, N. (2015). Multiple linear regression. *Nature Methods*, 12(12):1103–1104.

Lane, P. W. (2002). Regression analysis. In *Guide to GenStat release 6.1. Part 2. Statistics*. VSN International, Hemel Hempstead, England, United Kingdom.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology*, 36(3):214–224.

Lichtinghagen, R., Pietsch, D., Bantel, H., Manns, M. P., Brand, K., and Bahr, M. J. (2013). The enhanced liver fibrosis (ELF) score: normal values, influence factors and proposed cut-off values. *Journal of Hepatology*, 59(2):236–242.

Maiseli, B. J. (2019). Optimum design of chamfer masks using symmetric mean absolute percentage error. *EURASIP Journal on Image and Video Processing*, 2019(1):1–15.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529.

Makridakis, S. and Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure*, 405(2):442–451.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24.

Minitab Blog Editor (2013). Regression analysis: how do I interpret R-squared and assess the goodness-of-fit? `https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-` URL visited on 19th February 2021.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken, New Jersey, USA.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*,

78(3):691–692.
Nakagawa, S., Johnson, P. C., and Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213.
Nevitt, J. and Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education*, 68(3):251–268.
Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, 97(2):307.
Palechor, F.-M. and De-La-Hoz-Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, 25:104344.
Piepho, H.-P. (2019). A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical Journal*, 61(4):860–872.
Quinino, R. C., Reis, E. A., and Bessegato, L. F. (2013). Using the coefficient of determination. *Teaching Statistics: an International Journal for Teachers*, 35(2):84–88.
Raji, P. and Lakshmi, G. D. (2020). Covid-19 pandemic analysis using regression. *medRxiv*, 2020.10.08.20208991:1–8.
Rao, C. R. (1980). Some comments on the minimum mean square error as a criterion of estimation. Technical Report ADA093824, Pittsburgh University Institute for Statistics and Applications.
Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (2001). *Applied Regression Analysis: a Research Tool*. Springer Science & Business Media, Berlin, Germany.
Reeves, D. (28th January 2021). Personal communication (email).
Ren, L. and Glasure, Y. (2009). Applicability of the revised mean absolute percentage errors (MAPE) approach to some popular normal and non-normal independent time series. *International Advances in Economic Research*, 15(4):409–420.
Renaud, O. and Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140(7):1852–1862.
Rokach, L. and Maimon, O. (2005). Decision trees. In *Data Mining and Knowledge Discovery Handbook*, pages 165–192. Springer, Berlin, Germany.
Sammut, C. and Webb, G. I., editors (2010a). *Mean Absolute Error*, pages 652–652. Springer, Boston, Massachusetts, USA.
Sammut, C. and Webb, G. I., editors (2010b). *Mean Squared Error*, pages 653–653. Springer, Boston, Massachusetts, USA.
Sarbishei, O. and Radecka, K. (2011). Analysis of mean-square-error (MSE) for fixed-point FFT units. In *Proceedings of ISCAS 2011 – the 2011 IEEE International Symposium of Circuits and Systems*, pages 1732–1735. IEEE.
Saunders, L. J., Russell, R. A., and Crabb, D. P. (2012). The coefficient of determination: what determines a useful R^2 statistic? *Investigative Ophthalmology & Visual Science*, 53(11):6830–6832.
Seber, G. A. and Lee, A. J. (2012). *Linear Regression Analysis*, volume 329. John Wiley & Sons, Hoboken, New Jersey, USA.
Senapati, A., Nag, A., Mondal, A., and Maji, S. (2020). A novel framework for COVID-19 case prediction through piecewise regression in India. *International Journal of Information Technology*, 13(1):41–48.
Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., and Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, 24(24):171–176.
So, H. C., Chan, Y. T., Ho, K., and Chen, Y. (2013). Simple formulae for bias and mean square error computation. *IEEE Signal Processing Magazine*, 30(4):162–165.
Srivastava, A. K., Srivastava, V. K., and Ullah, A. (1995). The coefficient of determination and its adjusted version in linear regression models. *Econometric Reviews*, 14(2):229–240.
Sykes, A. O. (1993). An introduction to regression analysis. Law & Economics Working Papers 20, University of Chicago Law School Chicago Unbound.
Tötsch, N. and Hoffmann, D. (2021). Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer Science*, 7:e398.
University of California Irvine Machine Learning Repository (2019). Estimation of obesity levels based on eating habits and physical condition data set. `https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+`

physical+condition+ URL visited on 25th April 2021.
University of California Irvine Machine Learning Repository (2020). HCV data set. https://archive.ics.uci.edu/ml/datasets/HCV+data URL visited on 25th April 2021.
Wang, Q., Ma, Y., Zhao, K., and Tian, Y. (2020). A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26.
Wang, W. and Lu, Y. (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In *IOP Conference Series: Materials Science and Engineering*, volume 324, page 012049. IOP Publishing.
Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82.
Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, XX(7):557–585.
Wüthrich, M. V. (2020). From generalized linear models to neural networks, and back. Technical Report 3491790, RiskLab, Department of Mathematics, ETH Zürich.
Young, P. H. (2000). Generalized coefficient of determination. *Journal of Cost Analysis & Management*, 2(1):59–68.
Yue, Y. R. and Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, 55(1):84–96.
Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician*, 71(4):310–316.

SUPPLEMENTARY INFORMATION

		R^2	MAE	MSE	SMAPE	RMSE	MAPE
Random Forests	mean	**0.756**	0.149	0.133	1.808	0.361	0.092
Random Forests	s.d.	0.073	0.025	0.041	0.035	0.058	0.017
Linear Regression	mean	**0.535**	0.283	0.260	1.840	0.498	0.197
Linear Regression	s.d.	0.196	0.036	0.134	0.034	0.108	0.018
Decision Tree	mean	**0.423**	0.157	0.311	0.073	0.546	0.080
Decision Tree	s.d.	0.197	0.050	0.120	0.023	0.113	0.032

Table S1. Regression results on the prediction of hepatitis, cirrhosis, and fibrosis from electronic health records, including standard deviations. Mean values and standard deviations out of 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. We performed the analysis on the Lichtinghagen dataset (Lichtinghagen et al., 2013; Hoffmann et al., 2018) with the methods employed by Chicco and Jurman (2021). R^2: worst value $-\infty$ and best value +1. SMAPE: worst value 2 and best value 0. MAE, RMSE, MAPE, MSE: worst value $+\infty$ and best value 0.

		R^2	MAE	MSE	SMAPE	RMSE	MAPE
Random Forests	mean	**0.865**	0.412	0.512	0.087	0.714	0.094
Random Forests	s.d.	0.017	0.028	0.067	0.006	0.047	0.007
Decision Tree	mean	**0.426**	1.214	2.170	0.326	1.471	0.286
Decision Tree	s.d.	0.064	0.043	0.236	0.012	0.078	0.010
Linear Regression	mean	**0.254**	1.417	2.828	0.296	1.681	0.325
Linear Regression	s.d.	0.030	0.034	0.117	0.007	0.035	0.011

Table S2. Regression results on the prediction of obesity level from electronic health records, including standard deviations. Mean values and standard deviations out of 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. We performed the analysis on the Palechor dataset (Palechor and De-La-Hoz-Manotas, 2019; De-La-Hoz-Correa et al., 2019) with the methods Linear Regression, Decision Tree, and Random Forests. R^2: worst value $-\infty$ and best value +1. SMAPE: worst value 2 and best value 0. MAE, RMSE, MAPE, MSE: worst value $+\infty$ and best value 0.

j	R^2	cnSMAPE
5	0.9897	0.9500
6	0.9816	0.9400
7	0.9701	0.9300
8	0.9545	0.9200
9	0.9344	0.9100
10	0.9090	0.9000
11	0.8778	0.8900
12	0.8401	0.8800
13	0.7955	0.8700
14	0.7432	0.8600
15	0.6827	0.8500
16	0.6134	0.8400
17	0.5346	0.8300
18	0.4459	0.8200
19	0.3465	0.8100
20	0.2359	0.8000

Table 1. UC1 Use case. Values generated through Equation 11. R^2: coefficient of determination (Equation 3). cnSMAPE: complementary normalized SMAPE (Equation 10).

N	correct model		wrong model	
	R^2	cnSMAPE	R^2	cnSMAPE
2	–16.1555357	0.3419595	1	1
3	–0.1752271	0.5177952	1	1
4	0.7189524	0.6118408	1	1
5	0.7968514	0.6640983	1	1
6	0.8439391	0.7162407	1	1
7	0.8711581	0.7537107	1	1
8	0.8777521	0.7772273	1	1
9	0.9069923	0.7962306	1	1
10	0.9196087	0.8101526	1	1
11	0.9226216	0.8230926	-2.149735×10^{02}	0.9090909
12	0.9379797	0.8362582	-1.309188×10^{04}	0.8333333
13	0.9439415	0.8447007	-2.493881×10^{05}	0.7692308
14	0.9475888	0.8518829	-2.752456×10^{06}	0.7142857
15	0.9551004	0.8613108	-2.276742×10^{07}	0.6666667
16	0.9600758	0.8679611	-1.391877×10^{08}	0.6250000
17	0.9622725	0.8740207	-7.457966×10^{08}	0.5882353
18	0.9607997	0.8784127	-3.425546×10^{09}	0.5555556
19	0.9659541	0.8837482	-1.275171×10^{10}	0.5263158
20	0.9635534	0.8870441	-4.583919×10^{10}	0.5000000

Table 2. UC3 Use case. We define N, correct model, and wrong model in the UC3 Use case paragraph. R^2: coefficient of determination (Equation 3). cnSMAPE: complementary normalized SMAPE (Equation 10).

	R^2	MAE	MSE	SMAPE	RMSE	MAPE
Random Forests (RF)	**0.756**	0.149	0.133	1.808	0.361	0.092
Linear Regression (LR)	**0.535**	0.283	0.260	1.840	0.498	0.197
Decision Tree (DT)	**0.423**	0.157	0.311	0.073	0.546	0.080
rankings:						
1^{st}	RF	RF	RF	DT	RF	DT
2^{nd}	LR	DT	LR	RF	LR	RF
3^{rd}	DT	LR	DT	LR	DT	LR

Table 3. Regression results on the prediction of hepatitis, cirrhosis, and fibrosis from electronic health records, and corresponding rankings based on rates. We performed the analysis on the Lichtinghagen dataset (Lichtinghagen et al., 2013; Hoffmann et al., 2018) with the methods employed by Chicco and Jurman (2021). We report here the average values achieved by each method in 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. R^2: worst value -infinity and best value +1. SMAPE: worst value 2 and best value 0. MAE, MAPE, MSE, and RMSE: worst value +infinity and best value 0. We reported the complete regression results including the standard deviations in Table S1. R^2 formula: Equation 3. MAE formula: Equation 6. MAPE formula: Equation 7. MSE formula: Equation 4. RMSE formula: Equation 5. SMAPE formula: Equation 8.

method	R^2	MAE	MSE	SMAPE	RMSE	MAPE
Random Forests (RF)	**0.865**	0.412	0.512	0.087	0.714	0.094
Decision Tree (DT)	**0.426**	1.214	2.170	0.326	1.471	0.286
Linear Regression (LR)	**0.254**	1.417	2.828	0.296	1.681	0.325
rankings:						
1^{st}	RF	RF	RF	RF	RF	RF
2^{nd}	DT	DT	DT	LR	DT	DT
3^{rd}	LR	LR	LR	DT	LR	LR

Table 4. Regression results on the prediction of obesity level from electronic health records, including standard deviations. Mean values and standard deviations out of 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. We performed the analysis on the Palechor dataset (Palechor and DeLaHozManotas, 2019; DeLaHozCorrea et al., 2019) with the methods Linear Regression, Decision Tree, and Random Forests. We report here the average values achieved by each method in 100 executions with 80% randomly chosen data elements used for the training set and the remaining 20% used for the test set. R^2: worst value -infinity and best value +1. SMAPE: worst value 2 and best value 0. MAE, MAPE, MSE, and RMSE: worst value +infinity and best value 0. We reported the complete regression results including the standard deviations in Table S2. R^2 formula: Equation 3. MAE formula: Equation 6. MAPE formula: Equation 7. MSE formula: Equation 4. RMSE formula: Equation 5. SMAPE formula: Equation 8.