CRITIC ON R-SQUARE

Intepretation of R² as Surface Integral

Exposee

In this essay the author shows that R-Square is only applicable with proportional system behavior and very even distributed measurement data on the observation range.

Dr.-Ing. Wolfgang Rückert
info@ib-rueckert.de

Interpretation of coefficient of determinations' behavior for a linear and a non-linear case

In the literature there have been number of approaches for the interpretation of R^2. However, the author thinks it can add to the discussion to access the validity of R^2 by an interpretation of SS_{Tot} and SS_{Res} as numerical integrals.

Integrals as performance indicators

For optimization calculus, integrals are helpful to quantify problems that are extended in more than a point dimension. They can serve as an objective function or to qualitatively compare different approaches. The (squared) sums used in the coefficient of determination can be interpreted as numeric integrals. For the definition of R^2 see e.g. [1].

SS_{Tot} can be interpreted as the (squared) integral between measurement (output) data $y_{M,j}$ and a constant function $f_{const}(x)$, which represents the most simple case of a regression function with only one parameter constant over the observation range. Especially, the constant function corresponds to the mean value $\bar{y}$ of the measurement data

$$SS_{Tot} = \sum_j \left(y_{M,j} - \bar{y}\right)^2$$

SS_{Res} can be interpreted as the (squared) integral between the between measurement (output) data $y_{M,j}$ and a regression model of higher order.

$$SS_{Res} = \sum_j \left(y_{M,j} - f_{Regr}\left(\beta, x_{M,j}\right)\right)^2$$

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Tot}}$$

R^2 can be interpreted as an indicator that shows the improvement of regression by use of more elaborated models in comparison to the simplest possible model.

To check if the indicator R^2 is a valid indicator, two simple cases are investigated. The underlying relationship f_{true} be a line in the first case f_{line} and a parabola in the second case f_{parab}. It is assumed that the spread of measurement data is approaching zero and that all data points are equally distributed over the observation range. The cases are one dimensional, so that the integral yields a surface.

Minimum surface of a curve

There be a relationship $f_{true}(x)$ between input variable $x_{M,j}$ and an output variable $y_{M,j}$. The squared surface A of f_{true} on the observation range $[x_1, x_2]$ shall be defined with the help of a constant relationship

$$f_{const}(x) = \mu$$

There is a function that describes the distance between $f_{true}(x)$ and $f_{const}(x)$. The distance between the two functions is squared because the surface must not become negative when there are intersections.

$$f_{\Sigma} = (f_{true} - f_{const})^2$$

The surface (or volumes in case of more than a one dimensional input variable) is the integral of f_{Σ} on the conservation range.

$$A = \int_{x1}^{x2} f_{\Sigma} dx$$

The value of µ chosen freely results in an indefinite number of surfaces. There is only one characteristic value for the surface A, its minimum. Also, this would be the most consequential value regarding f_{const} as the most simple regression model. The value μ_{min} for the minimum surface A_{min} is found by

$$A = A_{min} \Leftrightarrow \frac{dA(\mu)}{d\mu} = 0$$

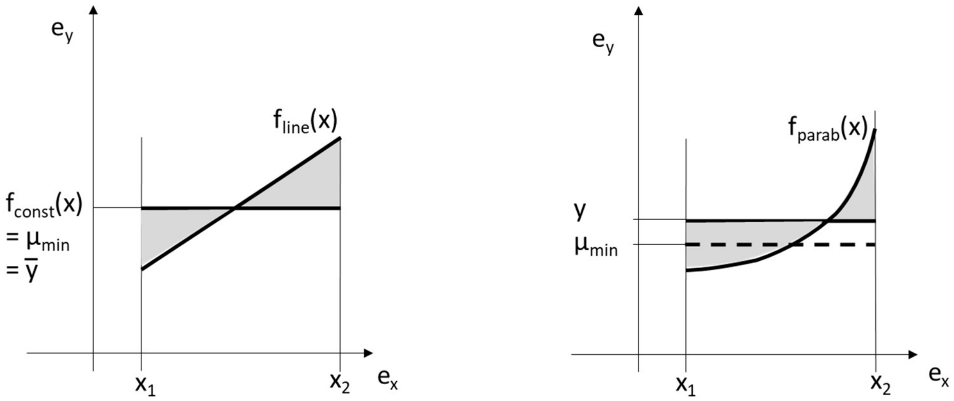

Figure 1 Minimum surface and curve mean for straight line and parabola

Mean value of a curve

Under the assumptions above (spread of measurement data is approaching zero and all data points are equally distributed over the observation range) the mean value of a bijective curve can be calculated as the center of gravity of the projection of the point density on e_y.

The point density $\varrho(y)$ is proportional to the slope of f_{true} with respect to e_y . The inverse function is needed.

$\varrho(y) = \frac{df_{true}^{-1}(y)}{dy}$

The center of gravity of $\varrho(y)$ is calculated with the integral on e_y.

$$\bar{y} = \int_{f_{true}(x1)}^{f_{true}(x2)} d\varrho(y)\, dy$$

Please note that the slope of the inverse of f_{true} does influence the calculated mean value in the sense of a weighing.

Comparison of results for line and parabola

In this paragraph we will examine whether μ_{min} and $\bar{y}$ are the same and discuss the implications.

For the straight line

$$f_{line}(x) = ax + b$$

$$\bar{y}_{line} = \frac{1}{2}(y_2 + y_1) = \mu_{min,line}$$

The straight line has a constant slope, so variances are weighed in even for the calculation of the mean.

For the parabola

$$f_{parab}(x) = ax^2$$

$$\bar{y}_{parab} = \frac{1}{2}a\,(x_2^2 - x_1^2)$$

$$\mu_{min,parab} = \frac{1}{3}a\,\frac{x_2^3 - x_1^3}{x_2 - x_1}$$

At least in one case other than a straight line, the mean value $\bar{y}$ of f_{true} corresponds not to the value for the minimal surface μ_{min}. If the measurement values are unevenly distributed on the observation range, the value of $\bar{y}$ gets even more aleatoric. This is because of the weighing with the inverse slope of f_{true}. The resulting SS_{Tot} will always been bigger then can be assumed for the best model f_{const}.

Discussion

From the mathematics above, one can see that the optimum value of r-square will never be 100% for at least one non-linear function (parabola without offset). The value of R-square depends strongly on the underlying “true” relationship and the distribution of the measurement points on the observation range due to weighing with the slope of the inverse of f_{true}. It therefore is not an objective indicator of the goodness of fit. The

Coefficent

R-Square is only

Applicable with

Proportional data Behavior.

Literature

[1] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang. Regression Modelle, Methoden und Anwendungen. Springer Berlin, Heidelberg. ISBN: 978-3-642-01837-4