Understanding the Metrics

Five metrics give us some hints about the *goodness-of-fit* of our model. The first two metrics, the *Mean Absolute Error* and the *Root Mean Squared Error* (also called *Standard Error of the Regression*), have the same unit as the original data. In fact, given *ŷ* the prediction, *y* the actual value and *n* the size of the sample, their definitions are the following:

$$e_i = y_i - \hat{y}_i$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

These metrics can be used to compare models having the error *e* measured in the same units. *MAE* is simpler to understand, since it describes the average error. *RMSE* is not as intuitive as the other metric. It should be used when large errors are not allowed, because they are squared and then averaged.

Both these metrics can range from 0 to ∞.

An interesting relation between them is given by the following inequations:

$$MAE \leq RMSE \leq MAE\sqrt{n}$$

So *RMSE* gets bigger than *MAE* as the sample size increases.

The other two metrics are the *Relative Absolute Error* and the *Relative Squared Error*, defined as following:

$$e_i = y_i - \hat{y}_i$$

$$RAE = \frac{\sum_{i=1}^{n} |e_i|}{}$$

$$\sum_{i=1} |y_i - y_i|$$

$$RSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y_i})^2}}$$

where *y bar* is the mean of the actual values *y*. They are derived from the first two metrics with the difference that there is a division by the variation of *y*. Because of that they are named "relative" and they can range from 0 to 1.

Since they are "relative", these metrics can be used to compare the accuracy between models having errors measured in different units.

All the above mentioned metrics (*MAE*, *RAE*, *RMSE*, *RSE*) are insensitive to the direction of errors (the signs of errors are removed by the absolute value and by squaring them). For all of these metrics lower values are better.

The last metric we have is the *Coefficient of Determination* or *R Squared* (*R²*). It is defined as following:

$$SS_{res} = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 \qquad \text{(Residual Sum of Squares)}$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y_i})^2 \qquad \text{(Total Sum of Squares)}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad \text{(Coefficient of Determination)}$$

The second addendum in the definition of *R²* can be seen as following:

$$\frac{SS_{res}}{SS_{tot}} = \frac{\frac{SS_{res}}{n}}{\frac{SS_{tot}}{n}} = \frac{VAR_{res}}{VAR_{tot}} = FVU$$

$$R^2 = 1 - FVU$$

where *FVU* is the *Fraction of Variance Unexplained*. The first equation compares the unexplained variance (variance of the model's residuals, *VARres*) with the total variance (of the actual values, *VARtot*). Since *FVU* ranges from 0 to 1 (*VARres*cannot be higher than *VARtot*), R^2 is what remains after subtracting the measure of unexplained (*FVU*) from the whole. So R^2 measures something related to the **explained variance**.

Taking a look at fig. 1, we have that $R^2 = 0.888678$. This means that 89% of the variability between the two variables has been accounted for (the "explained" part), and the remaining 11% of the variability is still unaccounted for. For wide classes of linear models, the balance equation between accounted and unaccounted variability can be expressed as following:

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 \qquad \text{(Explained Sum of Squares)}$$

$$SS_{tot} = SS_{res} + SS_{reg}$$

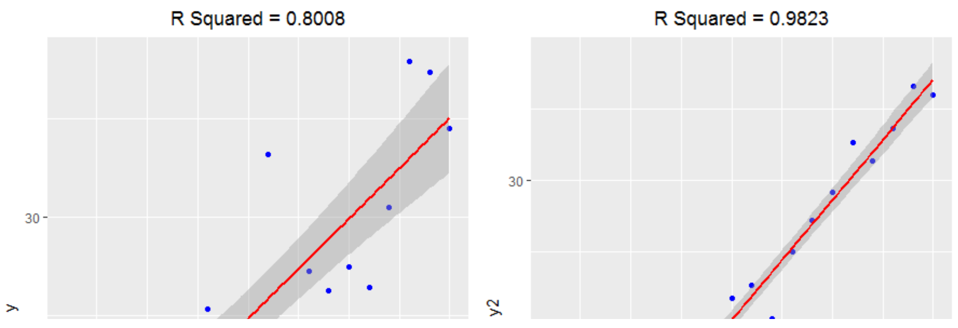
That said, we can rewrite the definition of R^2 as following:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

In this form:

> *R^2 can be seen as the percentage of the prediction variable's variation that is explained by a linear model.*

In other words, it's a measure of how close the data is to the fitted regression line:



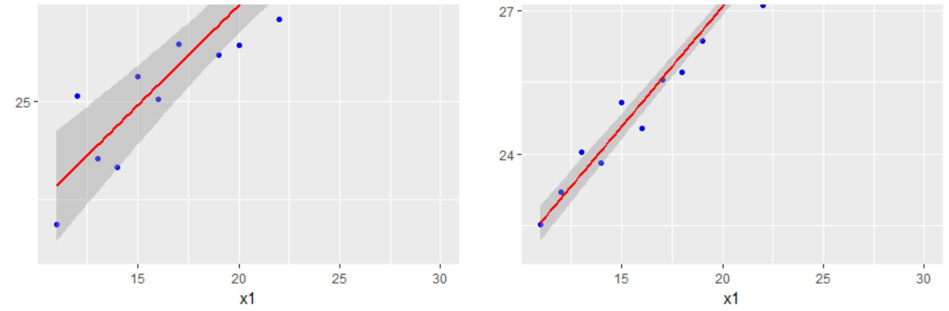


fig. 2—Evidence of the R² value in relation to the goodness-of-fitting

So if $R^2 = 0.888678$, then 89% of the total variation in y can be explained by the linear relationship between features and y.

This metric usually ranges from 0 to 1 and is unitless. But it can also be negative when the predictions are not obtained by linear regression. Contrary to other metrics, the closer R^2 is to 1, the better the model explains all the variability of the target variable around its mean.

The Right Way to Evaluate the Goodness-Of-Fit for Every Regression

Can R^2 tell us always the truth about the goodness-of-fit of our model? As you can imagine, the short answer is: no!

R^2 cannot tell us if the predictions are biased and sometimes it leads you to make bad decisions:

> R^2 can be low even if the model is good (the data contains an high amount of unexplainable variability)
>
> R^2 can be high even if the model is not good (the regressed function fits quite well but the resulting residuals are not randomly distributed, see later)

To make matters worst, all the assumptions done in the previous paragraph for R^2 take for granted the regression is **linear**.

> *If you are dealing with a nonlinear regression, R^2alone can lead to wrong conclusions. Only 28–43% of the models tuned*

> *using R^2 are correct.*

Specifically, for nonlinear regressions:

R^2 tends to be high for both very bad and very good models (even if you consider the *adjusted-R^2* defined later).

R^2 do not always increases for better nonlinear models.

In case of nonlinear regression, it's better to use the *Residual Standard Error* (or *Standard Error of Estimate* or *Standard Error of Regression*). It measures the average distance of the actual values from the regression line and it is conceptually similar to the standard deviation, with the difference that the standard deviation measures the average distance of the actual values from the mean. The *Residual Standard Error* (*S*) is defined below:

$$df = n - k - 1 \qquad \text{(Degrees Of Freedom of Residuals)}$$

$$S = \sqrt{\frac{SS_{res}}{df}} \qquad \text{(Residual Standard Error)}$$

where *k* refers to the number of predictors (parameters to be estimated using the regression), not including the intercept (it's accounted by the "- 1" in the *df* formula).