Useful and unuseful summaries of regression models

Roberto Todeschini
Milano Chemometrics and QSAR Research Group - Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)

In the scientific papers, it is very common to present the summary of a regression model in the following form:

$$\log P = 3.1347 - 0.0056 \cdot V1 + 12.567 \cdot V2$$
$$n = 15 \quad r = 0.9765 \quad s_y = 0.7612 \quad F = 18.1$$

where *logP* is the studied experimental response and V1 and V2 are two independent variables for which a quantitative relationship with the response has been searched for. The numbers in the equation are called ***regression coefficients***, being the first one the *intercept* of the regression model. *n* is the *number of samples* used for building the regression equation, *r* is the *multiple correlation coefficient*, s_y is the *residual standard deviation* (or error standard deviation), *F* the calculated value of the *F-ratio test*.
In several cases, $r = 0.9765$ is replaced by $R = 0.9765$ or by $R^2 = 0.9536$ or by $R^2 = 95.36$ %, being the last two indices (*coefficient of determination*) the squared *r* (*R*) value and the last one the corresponding percentage of the variance explained by the regression model.
For evaluating the quality of the summary, mathematical definitions of the quantities involved in the previous example are given.

- ***Residual Sum of Squares*** (*RSS*)

It is the sum of the squared difference between the experimental response y and the response calculated by the regression model:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \qquad (1)$$

If *RSS* is equal to zero the model is perfect, i.e. for all the *n* samples, the calculated responses coincide with the experimental responses. Obviously, *RSS* also depends on the measure unit used for the response. In practice, for the same model, if you multiply the experimental response for 10, *RSS* is 100 times greater, being a squared quantity.

- ***Total Sum of Squares*** (*TSS*)

It is the total variance that a regression model can explain and is used as a reference quantity to calculate standardized quality parameters. Also denoted as *SSY*, it is the sum of the squared differences between the experimental responses and the average experimental response:

$$TSS \equiv SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (2)$$

TSS is assumed as a theoretical reference model where for each experimental response a constant value is calculated as the average experimental response. As for *RSS*, also *TSS* depends on the measure unit used for the response.

- ***Derived regression parameters for evaluating the goodness of fit***

From the previous definitions of *RSS* and *TSS*, the following quantities are usually defined:

$$TSS = MSS + RSS \qquad (3)$$

where *TSS* and *RSS* are the quantities defined above and *MSS* is the ***Model Sum of Squares***.
All these three quantities are sums of squares and then are always positive quantities:

$$TSS \geq 0 \quad MSS \geq 0 \quad RSS \geq 0$$

The ***coefficient of determination*** R^2 and the ***multiple correlation coefficient*** *R* are defined as:

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad 0 \leq R^2 \leq 1 \qquad (4)$$

$$R \,(\text{or } r) = \sqrt{R^2} \qquad 0 \leq R \leq 1$$

The ***Mean square error*** (or *residual mean square*) s^2_y and the ***residual standard deviation*** (or *residual standard error)* s_y are defined as:

$$s_y^2 = \frac{RSS}{n - p'} \qquad s_y = \sqrt{\frac{RSS}{n - p'}} \qquad (5)$$

where *n* is the number of samples, *p*' the number of model parameters (often given by *p* variables plus the intercept). Other symbols that can be commonly used for the residual standard deviation are *RSD*, *SE*, and *s*.
Finally, the ***F-ratio test in regression*** is defined as the ratio between the variance explained by the model to the residual variance, both scaled by the corresponding degrees of freedom:

$$F = \frac{MSS / (p' - 1)}{RSS / (n - p')} \qquad (6)$$

It must be observed that, for the same number of objects and variables, if *RSS* decreases (i.e. a better model is obtained), then both R^2 and *F* increase monotonically with *RSS*, while the residual mean square s^2_y decreases monotonically with *RSS*. This means that the best goodness of fit can be equally evaluated by using any of the three parameters (max(R^2), max(*F*) and min(s^2_y).

Moreover, it must be noted that none of these parameters is related in any way to the model prediction power: they are all only related to the goodness of fit.

About r, R and R^2

First of all, it must be observed that, due to the mathematical properties of *R*, the goodness of fit of the following nested models

$$\log P = f(V1, V2) \qquad \text{(model 1)}$$
$$\log P = f(V1, V2, V3) \qquad \text{(model 2)}$$
$$\log P = f(V1, V2, V3, V4) \qquad \text{(model 3)}$$
$$\ldots\ldots$$
$$\log P = f(V1, V2, V3, V4, \ldots, V_p) \qquad \text{(model p)}$$

can always ranked as:

$$R(\text{mod}.p) \geq \ldots \geq R(\text{mod}.3) \geq R(\text{mod}.2) \geq R(\text{mod}.1)$$

i.e. the *R* value of a model A constituted by the same variables of another model B plus any variable is always greater than (or at least equal to) the *R* value of the model B.

This effect is due to the possible presence (and usually probable) of ***chance correlation***, i.e. even a completely unuseful variable (or a random variable) can enhance the *R* value, but never decreases it.

Therefore, when the aim is to select the best model in a population of models or select the best variables within a set of variables or compare different models, or get models containing only relevant variables, the parameter R cannot be used (neither r or R^2, obviously).

In order to overcome these drawbacks, ***Adjusted R^2*** (R^2_{ADJ}) can be used in place of R^2, since increasing number of unnecessary variables gives a penalty to the R^2. It is defined as:

$$R^2_{ADJ} = 1 - (1 - R^2) \cdot \left(\frac{n-1}{n-p} \right) \qquad (7)$$

Note that *Adjusted* R^2 coincides with R^2 for model containing just one independent variable. However, remember that even adjusted R^2 is not related in any way to the real model prediction ability.

About the error standard deviation

The error standard deviation s_y is a statistical estimate of the error also accounting for the model degrees of freedom. This quantity is in the same unit of the response and is based on the fitting performance of the model, i.e. *RSS*. A more direct measure of the average error of the response estimates is the ***Standard Deviation Error in Calculation*** (*SDEC* or *SEC*):

$$SDEC \equiv SEC = \sqrt{\frac{RSS}{n}} \qquad (8)$$

where *n* is the number of samples.

About the F-ratio test

In the few last decades, the majority of regression model summaries include the *F* value. Analogously to adjusted R^2, higher the *F* value better the model. This use of the *F* value in evaluating a regression model is correct, but some problems arise when one think that the F value was originally proposed as a statistical test, that is the calculated F needs to be compared with a critical value at some probability level in order to draw decisions.

In effect, the null hypothesis H_0 of the *F*-ratio test in regression states that *all* the regression coefficients are equal to zero (i.e. no regression model is obtained) against the alternative hypothesis H_1 that *at least one* of the regression coefficients is different from zero (i.e. a regression model is obtained).

Therefore, in multivariate analysis, where more than one independent variable is used, this test is very weak and, in practice, completely unuseful: in fact, we want *all the variables* included in the model to be relevant (or statistically significant) for the response we are modelling.

Regression parameters related to the prediction power (goodness of prediction)

In statistics and chemometrics several ***validation techniques*** have been proposed in the last 20 years in order to estimate the model prediction capability. The model prediction capability is something different from the model fitness capability, i.e. the ability of the model to estimate the response for objects that do not participate to the calculated model.

Here, the attention is stressed on the most simple way to obtain some measures of model ability, i.e. the ***leave-one-out cross-validation technique***. This technique is often too optimistic with respect to the "true" prediction power of the models; however, it can be easily carried out and allows model comparison as well as variable selection.

The measures of model predictive ability are very similar to those defined for the goodness of fit and are based on the Predictive Error Sum of Squares (*PRESS*) which substitutes the Residual Sum of Squares (*RSS*):

- **Predictive Error Sum of Squares** (*PRESS*)

It is the sum of the squared differences between the experimental response *y* and the response *predicted* by the regression model, i.e. for an object that was not used for model estimation.

It is defined as:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i/i} \right)^2 \qquad (9)$$

where the notation i/i indicates that the response is predicted by a model estimated when the i-th sample was left out from the training set.

Therefore, an equivalent parameter to R^2 can be defined, using in place of *RSS* the quantity *PRESS*. This parameter is called cross-validated R^2 and the accepted symbols are R^2_{CV} or Q^2 :

$$R^2_{CV} \equiv Q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad Q^2 \leq 1 \qquad (10)$$

For bad predictive models, Q^2 can assume even negative values when *PRESS* is greater than *TSS*, meaning that in prediction the model performs worse than the no-model estimate, i.e. the mean response of the training set.

Note. A discussion about different Q^2 functions was recently published in the references given below, where a modified general Q^2 function is also proposed.

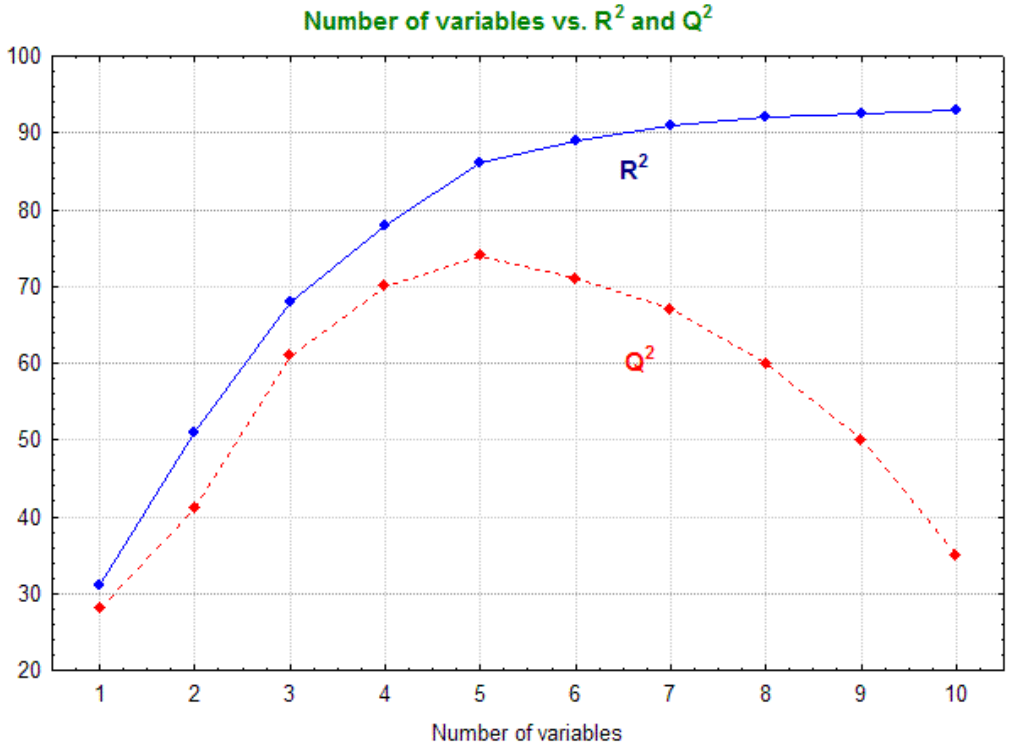
1) Consonni, V., Ballabio, D. and Todeschini, R. (2009) Comments on the definition of the Q^2 parameter for QSAR validation. *Journal of Chemical Information and Modeling*, **49**, 1669-1678.

2) Consonni, V., Ballabio, D. and Todeschini, R. (2010) Evaluation of model predictive ability by external validation techniques. *J. Chemometrics*, **24**, 194-201.

The **predictive squared error** (*PSE*) and the **standard deviation error in prediction** (*SDEP or SEP*) on the modeled response are given by:

$$PSE = \frac{PRESS}{n} \qquad SDEP \equiv SEP = \sqrt{\frac{PRESS}{n}} \qquad (11)$$

The characteristic behaviour of R^2 and Q^2 is shown in the figure below.



As it can be noted, R^2 values always increase when one variable at-a-time is added to the previous model variables; on the other side, Q^2 values increase until useful variables are added to the

previous model variables; when noisy variables are added to the model, Q^2 values decrease, i.e. the prediction power of the model is lowered. Therefore, the optimal complexity of the model – the best model predictive power - can be chosen by simply looking at the maximum value of Q^2 (obviously, the same model ranking is obtained looking at the minimum values of *PSE* or *SDEP*).

Examples of some questionable regression summaries

In order to better explain the different meanings of the parameters used for a summary of regression models, some examples taken from literature are given below.

Example 1 – Comparison / selection of regression models

In a paper, five different models have been calculated for modelling binding affinities of 49 compounds, using five different sets of variables, each of ten different descriptors. The summary of the five models is collected in the Table below. *Here* Q *is the so-called "quality index" and not the cross-validated Q parameter* (compare equations 10 and 12 and example 2 given below).

1.	$n = 49$	$r = 0.880$	$s_y = 0.451$	$F = 13.05$	$Q = 1.951$	$R^2 = 0.774$	$R^2_{CV} = 0.672$
2.	$n = 49$	$r = 0.932$	$s_y = 0.308$	$F = 25.56$	$Q = 3.026$	$R^2 = 0.868$	$R^2_{CV} = 0.802$
3.	$n = 49$	$r = 0.914$	$s_y = 0.423$	$F = 19.32$	$Q = 2.160$	$R^2 = 0.836$	$R^2_{CV} = 0.737$
4.	$n = 49$	$r = 0.938$	$s_y = 0.319$	$F = 28.45$	$Q = 2.940$	$R^2 = 0.880$	$R^2_{CV} = 0.814$
5.	$n = 49$	$r = 0.942$	$s_y = 0.412$	$F = 24.40$	$Q = 2.286$	$R^2 = 0.887$	$R^2_{CV} = 0.795$

First of all, it there is redundant information due to the presence of both r and R^2 values. Moreover, having all the models the same number of samples (49) and model parameters (10+intercept), the degrees of freedom for the numerator and denominator of the F-ratio are 10 and 38, respectively, with a F critical value (at 5%, 1% and 0.1% of confidence) of 2.09, 2.81 and 3.90, respectively: using the F test, all the models are more than acceptable at any level of confidence (see also the above paragraph *About the F-ratio test*).

On the basis of R^2 (and r), the models can be ranked (from the best to the worst) with the following order: 5, 4, 2, 3, 1. The same order *should be* obtained by ranking the models on the basis of s_y (lower the value, better the model) and by using the quality index Q (see example 2). As for R^2, also these two parameters depend on *RSS*, *TSS* being a constant, provided that the measure unit of the response was not changed from one model to another. However, the model ranking is: 2, 4, 5, 3, 1 in both cases, thus indicating the presence of some mistakes in reporting the results or in the calculations. Disregarding the previous wrong rankings, the unique reasonable model ranking is that obtained by R^2_{CV}: 4, 2, 5, 3, 1, i.e. evaluating the model predictive power.

The Authors claim that model 5 is better than model 4 and that it is the best one because "Introduction of the new ... [index] ... in QSAR model for σ receptor increased the predictive value of the model.": this sentence is false.

Example 2 - The quality index

Another interesting model, reported in literature, is the following:

$$\log P = 4.1639 + 0.0048 \cdot W$$
$$n = 16 \quad r = 0.9703 \quad s_y = 0.8434 \quad Q = 1.1505$$

In 1994, a quality index Q for regression was defined as:

$$Q = \frac{R}{s_y} \qquad (12)$$

where R and s_y are the measures of goodness of fit defined in equations (4) and (5), respectively. Several people have used this quality index without any criticism (see also example 1). However, it can be easily demonstrated that it is based on a very doubtful statistics. In effect,

$$Q^2 = \frac{R^2}{s_y^2} = \frac{\frac{MSS}{TSS}}{\frac{RSS}{(n-p')}} = \frac{MSS}{RSS} \cdot \frac{(n-p')}{TSS} = F \cdot \frac{(p'-1)}{(n-p')} \cdot \frac{(n-p')}{TSS} = F \cdot \frac{(p'-1)}{TSS} \qquad Q = \sqrt{\frac{F'}{TSS}}$$

where F' is

$$F' = \frac{MSS / (p'-1)}{RSS / (n-p')} \cdot (p'-1) = \frac{MSS / 1}{RSS / (n-p')}$$

Therefore, the Q index is related to a modified F test, which takes into account only the degrees of freedom of the model residual sum of squares.

Moreover, Q also depends on the total sum of squares TSS, which is constant for a given response. However, if the measure unit of the response is changed (for example, multiplied by 10), the TSS value will increase 100 times and the quality decreases consequently. On the contrary, if the response is divided by 10, the quality increases proportionally.

In fact, if the response is defined as multiplied by a scaling factor c, the total sum of squares is:

$$TSS = \sum_{i=1}^{n} (c \cdot y_i - c \cdot \bar{y})^2 = c^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (xx)$$

Therefore, the quality index Q can be also expressed in a simpler form as:

$$Q = \frac{1}{c} \cdot \sqrt{\frac{F'}{TSS}}$$

$$Q^2 = \frac{R^2}{RSS / (n-p')} = \frac{R^2}{(1-R^2)} \cdot \frac{(n-p')}{TSS} = \frac{R^2}{(1-R^2)} \cdot \frac{(n-p')}{c^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

thus, showing that, for *n* objects and *p* variables, Q increases monotonically (but not linearly) with *R* (as well as with the decreasing *of* s_y), being *TSS* constant. However, if a change in the response scale is performed, the quality changes as 1/*c*.

But, in the case of the model shown above, another problem arises: in fact, the Authors confound the quality index Q with the cross-validated *Q* (or Q^2 or R^2_{CV}) parameter, then attributing to the quality index predictive properties which on the contrary are completely absent:

"This quality factor Q is defined as the ratio of the correlation coefficient (*r*) to the standard deviation (Sy) i.e., Q = r/S_y. This factor accounts for the predictive power of the model." .

As it can be easily observed, none of the parameters in the quality index definition is in some way related to the prediction power of the model, but is (of course) related to *R*.

Example 3 - Negative r values

Look at the model summary below:

$$\log k_i = 4.0539 - 0.0018 \cdot Sz$$
$$n = 16 \quad r = -0.5809 \quad s_y = 0.4438 \quad F = 7.130$$

In this model, *r* (the multiple correlation coefficient) is presented with the minus sign, due to the inverse pairwise correlation between the variables logk_i (the response) and *Sz* (the descriptor). The reported *r* is wrong, being the multiple correlation coefficient (see equations 3 and 4) always positive because it represents (or is related to) the explained variance of the model. The correct expression is *r* = 0.5809, while the information related to the existing inverse correlation between logk_i and *Sz* is well highlighted by the sign of the regression coefficient.

This is another error which sometimes appears in using *r* or *R* due to the counfounding of the multiple correlation coefficient (just *r* or *R*) with the bivariate correlation *r*, which represents the correlation between a pair of variables and ranges between –1 and +1, defined as:

$$r(x,y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \qquad (13)$$

Of course, for bivariate models, the absolute value of *r*(*x*,*y*) coincides with *R*.

A suggested summary of a regression model

From the above considerations, it seems quite reasonable to characterize the models with a simple but effective summary, separating the prediction and the fitness qualities of the models. Moreover, in order to avoid confusions, it is suggested to use capital *R* in place of *r*, leaving to the latter the meaning of pairwise correlation between two variables. Moreover, the square values of both *R* and *Q* give an explicit idea of the modeled variance. Once the model has been accepted, a better

information about the average error (in fitting and prediction) on the modeled response is given by the parameters *SDEC* and *SDEP*. Finally, in order to avoid pity discussions with some referees, sometimes the *F*-ratio test can be also included, giving the degrees of freedom and the corresponding probability.

$$\log P = 3.1347 - 0.0056 \cdot V1 + 12.567 \cdot V2$$
$$n = 15 \quad Q^2_{LOO} = 93.62\% \quad SDEP = 0.792$$
$$R^2 = 97.65\% \quad SDEC = 0.821$$

Obviously, other estimates of the prediction power performed on an *external test set* (reporting also the number of objects in the test set) or using *leave-many-out* (together with the percentage of objects left out in each step) or *boostrap* or other validation techniques are welcome and should be also reported, together with other specific information about the adopted validation technique.

$$n = 15 \quad Q^2_{LMO}(20\%) = 91.13\% \quad Q^2_{LMO}(30\%) = 90.32\% \quad Q^2_{BOOT} = 90.56\%$$
$$n = 5 \quad Q^2_{EXT} = 88.03 \quad SDEP_{EXT} = 0.872$$