TONY DAVIES COLUMN

Back to basics: calibration statistics

A.M.C. Davies[a] and Tom Fearn[b]
[a]Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK; [b]Department of Statistical Science, University College London, Gower Street, London, UK

Introduction

It had been intended that this column would be about PLS calibration, but before we do this we need to fill a gap. When we compute calibrations we need some "figure of merit" to decide if we have a "good" calibration. These are calibration statistics which have not been discussed in this column for a long time.

r^2?

Many people know and use a calibration statistic—the square of the product–moment correlation coefficient, r^2—so why not just use this? There is a problem with r^2 but a demonstration may be more persuasive than a lecture.

Figure 1 is a scatter plot of a calibration. It is actually a simulated calibration using data calculated from the model:

$$y=a+bx+e$$

with the "*e*"s normally distributed and in the data shown $a=0$ and $b=1$. The calibration was produced by a calibration set of 200 samples with a range of values from 10 to 20 and a variability that gives it an r^2 of 0.903.

Figure 2 shows a test of the calibration with a test set of 100 samples with a range of values from 13 to 17. The r^2 was 0.625. Judging by the r^2 this is not a good calibration. If we increase the range of the test set to be between the values of 12 and 18 the r^2 rises to 0.786, Figure 3. If we increase the range of the test set to match the calibration (values of 10 to 20), Figure 4, then the r^2 increases to 0.911. The value of r^2 depends on the range!

SEP

If r^2 is not the best choice for measuring the merit of a calibration, then what is? The answer is the standard error of performance (or prediction), *SEP*, which is a measure of the variability of the difference between the predicted and reference values for a set of validation samples. The starting point is a set of reference values r_i and of predictions p_i for n samples in a validation set. Let $d_i=r_i-p_i$ be the difference between reference and predicted for the ith sample. Then the simplest way of calculating *SEP* is as the root mean square of differences:

$$RMSEP=\sqrt{\{(\Sigma d_i^2)/n\}}$$

where the sum, like all those that follow, is over all samples, i.e. from 1 to n. Because all the differences get squared, it does not matter whether one uses r_i-p_i or p_i-r_i in the calculation of d_i. This *SEP* is an estimate of what a typical difference between prediction and reference values is likely to be when the calibration is used for real. As such, it rightly

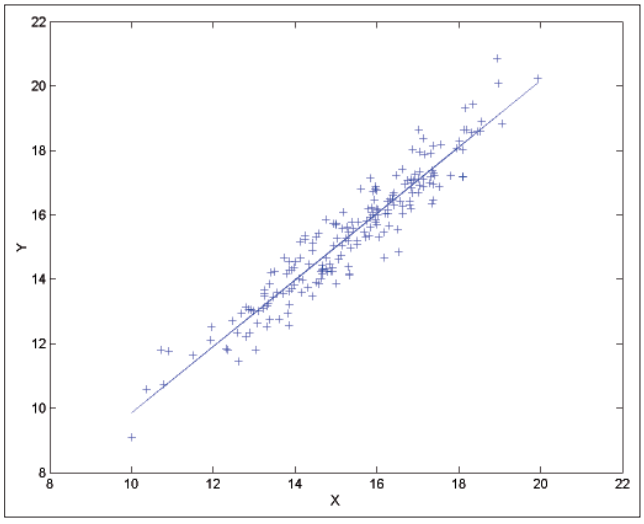

Figure 1. Plot of the simulated calibration. $r^2=0.90$.

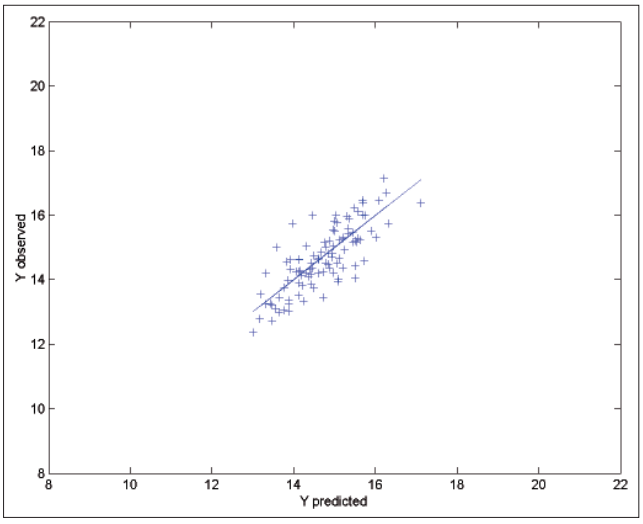

Figure 2. Plot of results from a test set of 100 samples in the range 13–17; $r^2=0.625$.

TONY DAVIES COLUMN

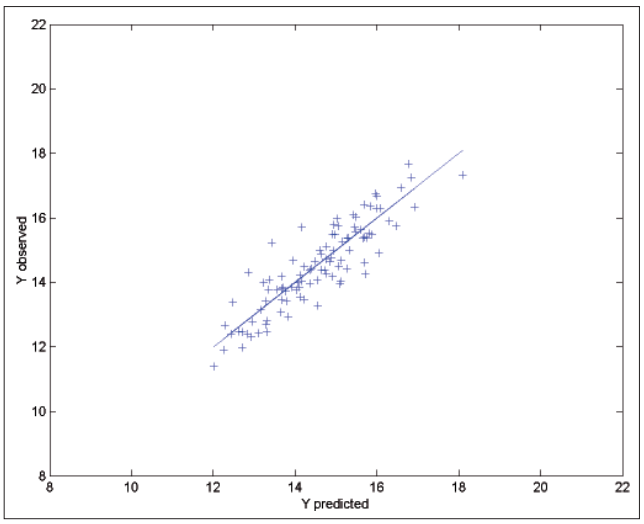

Figure 3. Plot of results from a test set of 100 samples in the range 12–18; r^2=0.786.

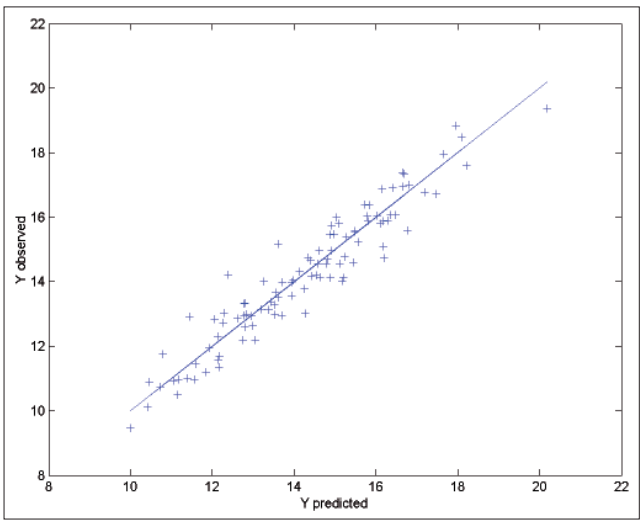

Figure 4. Plot of results from a test set of 100 samples in the range 10–20; r^2=0.911.

includes contributions from several sources, including errors in the reference measurements.

Unfortunately there are other versions of the calculation, the most common being the one which is corrected for bias, SEP_{-b}.

First calculate the average difference

$$bias = (\Sigma d_i)/n$$

Then subtract this from each difference in the calculation of SEP_{-b}, giving

$$SEP_{-b} = \sqrt{[\{\Sigma(d_i - bias)^2\}/(n-1)]}$$

It can be shown that* the relationship between the two is:

$$RMSEP^2 = SEP^2_{-b} + bias^2$$

If we were able to know what the bias would be when we predict real samples, SEP_{-b} would be a useful statistic but if the *bias* has changed since the calibration was done then there is no way of knowing that it will not change again. The tendency is for people to forget that they have removed the bias! So either use *RMSEP* or *bias* **and** SEP_{-b}. If the bias is small, then *RMSEP* and SEP_{-b} will be similar. The values of these statistics for the previous examples are shown in Table 1.

Table 1. Statistics for the examples shown in Figures 2–4.

Variable	Test set 1	Test set 2	Test set 3
Range	13–17	12–18	10–20
r^2	0.625	0.786	0.911
RMSEP	0.615	0.615	0.616
SEP_{-b}	0.616	0.616	0.618
bias	–0.050	–0.045	–0.036

With a simple linear regression like this one, the *SEC*, calculated from the differences between reference and predicted values for the training data by a formula like that for *RMSEP* but with a divisor of $n-2$ to allow for the fact the prediction line has been fitted to these data, is a reliable predictor of what the *SEP* should be on validation data. Here the value of 0.621 is indeed close to the observed *SEP*s. Unfortunately, in the case of multivariate calibration with large numbers of predictors, *SEC* is not a reliable guide to future performance, and we need a validation set to estimate *SEP* directly.

Conclusion

Chemometric software will probably continue to compute and display r^2 values for ever and if it is there it is impossible not to take note of it, but be guided by the *RMSEP* or SEP_{-b} **and** *bias* values. Calibrations should, of course, be tested over the range in which they will be used.

---

*For readers who enjoy some algebra!

Given that

$$RMSEP = \sqrt{\left[\left(\sum_{i=1}^{n} d_i^2\right)/n\right]}$$

or $n\,RMSEP^2 = \sum_{i=1}^{n} d_i^2$ and $bias = (\sum_{i=1}^{n} d_i)/n$

or $n\,bias = \sum_{i=1}^{n} d_i$

and $SEP_{-b} = \sqrt{(\sum_{i=1}^{n}(d_i - bias)^2)/(n-1)}$

Square and multiply by $(n-1)$

$$(n-1)SEP^2_{-b} = \sum_{i=1}^{n}(d_i - bias)^2$$

Expand

$$(n-1)SEP^2_{-b} = \sum_{i=1}^{n}(d_i - 2bias\,d_i + bias^2)$$

Sum individual terms

$$(n-1)SEP^2_{-b} = \sum_{i=1}^{n} d_i^2 - 2bias\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} bias^2$$

Substitute for

$$\sum d_i^2 = n\,RMSEP^2 \text{ and } \sum_{i=1}^{n} d_i = n\,bias$$

note that $\sum_{i=1}^{n} bias^2 = n\,bias^2$

$(n-1)\,SEP^2_{-b} = n\,RMSEP^2 - 2n\,bias^2 + n\,bias^2$

Thus, ignoring the difference between n and $n-1$ and re-arranging

$$RMSEP^2 = SEP^2_{-b} + bias^2$$