Chemometrics and Intelligent Laboratory Systems 105 (2011) 220–222

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Short Communication

Using R^2 to compare least-squares fit models: When it must fail

Joel Tellinghuisen [a,*], Carl H. Bolster [b]

[a] Department of Chemistry, Vanderbilt University, Nashville, TN 37235, United States
[b] U. S. Department of Agriculture, Agricultural Research Service, 230 Bennett Lane, Bowling Green, KY 42104, United States

ARTICLE INFO

Article history:
Received 9 November 2010
Received in revised form 4 January 2011
Accepted 5 January 2011
Available online 13 January 2011

Keywords:
R^2
Data transformation
Weighted least squares
Model comparison
Michaelis–Menten
Langmuir

ABSTRACT

R^2 can be used correctly to select from among competing least-squares fit models when the data are fitted in common form and with common weighting. However, when models are compared by fitting data that have been mathematically transformed in different ways, R^2 is a flawed statistic, even when the data are properly weighted in accord with the transformations. The reason is that in its most commonly used form, R^2 *can* be expressed in terms of the excess variance (s^2) and the total variance in y (s_y^2) — the first of which is either invariant or approximately so with proper weighting, but the second of which can vary substantially in data transformations. When given data are analyzed "as is" with different models and fixed weights, s_y^2 remains constant and R^2 is a valid statistic. However, then s^2, and χ^2 in weighted fitting, are arguably better metrics for such comparisons.

1. Introduction

R^2 is probably the most familiar metric for judging goodness of fit — disparaged by many but still used routinely by many more. Its shortcomings have inspired many comments and suggestions for improvement [1–9]. Its well-known misuses include in the selection of a fit model when differently transformed data are fitted without attention to the data weighting changes that should accompany such transformations [10,11]. However, it turns out that it fails in this situation even when data weighting is properly taken into account. The reason for this failure is quite simple, but to our knowledge has not been given before now.

In its most widely used form, R^2 is the square of the coefficient of correlation (Pearson's R) between x and y for a set of n points (x_i,y_i),

$$R = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2(y_i-\bar{y})^2}} \equiv \frac{S_{XY}}{\sqrt{S_{XX}\times S_{YY}}} \quad (1)$$

where overbars designate averages. R is also given by $R=(b\ b')^{1/2}$, where b and b' are the least-squares (LS) slopes for linear regression of y upon x and x upon y, respectively; and by

$$R^2 = 1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2} \equiv 1-\frac{SSE}{S_{YY}} \quad (2)$$

where $\hat{y}_i$ is the calculated value of y at x_i from the regression of y upon x. In the latter form, the second term is recognized as the ratio of the residual sum of squares to the total sum of squares, and it expresses the extent to which the fit model accounts for the variability in y. Accordingly, R^2 is sometimes said to represent the efficiency of the LS fit. A perfect fit yields $R^2=1$. These definitions are for homoscedastic data (constant σ) but are extended to heteroscedastic data by incorporating a weight w_i in each summation, with averages now representing weighted averages [12]. If the weights are taken as the inverse variances, $w_i=\sigma_i^{-2}$, as required for minimum variance estimation, *SSE* is an estimate of χ^2 for the fit [13].

Eq. (2) and its weighted version are commonly used also for both linear and nonlinear LS fits to models other than the straight line, including in widely used data analysis programs (*e.g.*, SAS and KaleidaGraph). It is in such applications that this definition of R^2 has come in for most of its criticism. However, with care it can still be used in this form, in which it is essentially equivalent to χ^2 (s^2 for unweighted fits of homoscedastic data), though arguably harder to "read". On the other hand, it cannot be used at all for comparing LS fits involving differently transformed data, which seems to have been first noted by Scott and Wild [4], but without explanation.

To see why R^2 fails in this situation, consider different mathematically equivalent forms of a given fit model. A prime example, and the one that brought us to this awareness, is the rectangular hyperbolic relation,

$$y = \frac{abx}{1+bx} \quad (3)$$

of importance in kinetics work (Michaelis–Menten equation), in sorption (Langmuir), in binding and complexation, and in

* Corresponding author.
E-mail address: joel.tellinghuisen@vanderbilt.edu (J. Tellinghuisen).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.01.004

fluorescence quenching. Eq. (3) can be linearized a number of ways, including

$$Y \equiv \frac{1}{y} = \frac{1}{abx} + \frac{1}{a} \equiv CX + A \tag{4}$$

$$Y \equiv \frac{x}{y} = \frac{1}{ab} + \frac{x}{a} \equiv C + Ax \tag{5}$$

and

$$Y \equiv \frac{y}{x} = ab - by. \tag{6}$$

All of these transformations alter the data weighting [14], since, by error propagation, $\sigma_{Y4} = \sigma_y/y^2$, $\sigma_{Y5} = x\sigma_y/y^2$, and $\sigma_{Y6} = \sigma_y/x$; in addition Eq. (6) requires a correlated treatment, since the dependent variable is taken to be a function of itself [15]. When the fits to Eqs. (3–6) are carried out with proper weighting, they yield comparable *SSE* values — or identical values in LS algorithms based on Deming's [16] treatment [15,17,18]. But S_{YY} does *not* remain constant under these transformations, making R^2 an unreliable metric that can incorrectly prefer one form of Eqs. (3)–(6) over the others. It follows that fitting transformed data to mathematically different fit models may also yield different R^2 for identical *SSE* values.

2. Illustrations

Table 1 presents two 8-point synthetic data sets for the model of Eq. (3), generated for a roughly geometric *x*-structure. The relevant LS statistical properties were evaluated for these two data sets — for constant data error and for proportional error in *y* (constant coefficient of variation) — and for exact data. As summarized in Table 2, S_{YY} for the exact data varies considerably with choice of Eqs. (3)–(5), and the R^2 test would thus prefer Eq. (5) for homoscedastic data and Eq. (3) for proportional error in *y*. These predictions are borne out in the results for the two synthetic data sets, which incidentally give identical *SSE* values for Eqs. (4) and (5) for both error structures, even for conventional nonlinear LS algorithms [15]. (For a given error structure, all four forms yield the Eq. (3) results in the Deming-based algorithm). Note that the values for R^2 are independent of the scale of the weights, which were assigned from the standpoint of an analyst who knows the error structure (constant or proportional) but not its scale. Both *SSE* and S_{YY} scale with the weights, and the scale factor must be preserved on transformation to make the *SSE* comparisons meaningful.

For unweighted fitting, the S_{YY} values for the exact model in Table 2 do not tell the whole story, because the values of *SSE* also change. To check this dependence, we ran Monte Carlo simulations on synthetic data sets having the two error structures. These computations employed methods like those described previously [18]. We see that Eq. (5) wins the R^2 competition for both error structures. These results are consistent with the widespread preference for Eq. (5) in soil sorption work, often made on just such comparisons [10]. By contrast, the statistically best results are obtained for the version having true weights most nearly constant (which is the assumption behind unweighted LS). That is Eq. (3) for constant σ and Eqs. (3) and (4) coequally for proportional error; yet Eq. (3) failed to win a single time in the latter case. (The actual situation in soil sorption is more complex, involving data variance functions that are effectively intermediate between the limiting models considered here [19].)

Table 1
Synthetic data in the model of Eq. (3), having $a = 5$, $b = 0.2$, and two error structures.

x	*y*[a]	
	$\sigma = 0.2$	$\sigma = 0.08 \times y$
0.5	0.62	0.49
1	1.06	1.01
2	1.58	1.47
5	2.37	2.44
10	3.14	3.43
20	3.96	3.92
50	4.68	4.17
100	4.33	5.24

[a] Data obtained by adding random normal error scaled by the indicated σ values, to the true curve of Eq. (3).

Table 2
Properties of model of Eqs. (3)–(5) under various treatments.

	Equation		
	3	4	5
S_{YY}, exact model[a]			
Unweighted	19.92	3.421	343.5
$\sigma = 0.2$	498.1	89.32	910.5
$\sigma = 0.08 \times y$	596.1	312.7	407.8
synthetic, $\sigma = 0.2$[b]			
SSE	0.2602	0.2745	0.2745
S_{YY}	16.74	3.894	39.42
R^2	0.9845	0.9295	0.9930
Synthetic, $\sigma = 0.08 \times y$[c]			
SSE	0.03452	0.03815	0.03815
S_{YY}	3.682	1.958	2.832
R^2	0.9906	0.9805	0.9865
Unweighted, Monte Carlo[d]			
$\sigma = 0.2$	5.9%	3.3%	90.8%
$\sigma = 0.08 \times y$	0	39.6%	60.4%

[a] $w_i = 1$ for all unweighted; inverse variance weighting for other two rows. All *SSE* values obtained by minimizing the sum of weighted squared residuals for the respectively defined dependent variables.
[b] $w_i = 1$ for Eq. (3), y_i^4 for Eq. (4), y_i^4/x_i^2 for Eq. (5).
[c] $w_i = y_i^{-2}$ for Eq. (3), y_i^2 for Eq. (4), y_i^2/x_i^2 for Eq. (5).
[d] from 10^4 data sets for each error structure. Percent yielding highest R^2 for indicated equation; all fitting unweighted.

3. Valid use of R^2

As has been noted, R^2 comparisons remain valid when data are fitted to models without transformation. First consider, unweighted fits. The assumption is that the data are homoscedastic of unknown σ. Assuming a fit to a true model, Eq. (2) can be expressed as

$$R^2 = 1 - \frac{(n-p)s^2}{(n-1)S^2} \tag{7}$$

where *p* is the number of adjustable parameters, s^2 is an estimate of σ^2, and S^2 is the total variance in *y*, equivalent to the estimated variance for a fit to the model $y = a$ (yielding for *a* the average of *y*). It follows that as long as the model includes a constant, *SSE* can be no greater than S_{YY}, ensuring that $R^2 \geq 0$ [5]. Since LS minimizes *SSE*, in a test of different fit models having the same *p*, maximizing R^2 is the same as minimizing s^2. For comparing models of different *p*, an obvious problem occurs, since adding parameters can only decrease *SSE*. To counter this, an adjusted R^2 has been recommended [1,5],

$$R^2_{adj} = 1 - \frac{(n-1)SSE}{(n-p)S_{YY}} = 1 - \frac{s^2}{S^2} \tag{8}$$

which, in effect, preserves the equivalence between maximizing R^2 (now R^2_{adj}) and minimizing s^2.

A problem with this usage of R^2 is that it can compress all the variability inherent in s^2 into a very small range of values close to unity [12]. For example analytical chemists look for "three nines" in calibration fits of data obtained in many techniques. For the sake of illustration, suppose $s^2 = 1$, $S^2 = 1000$, and $n = 10$. Since s^2 is χ^2-distributed, it has relative standard deviation $(2/\nu)^{1/2} = 1/2$, where ν is

the number of degrees of freedom ($=n-p$, with $p=2$ for the linear response function). Thus one can expect values of s^2 in the approximate range 0.5–1.5 about 2/3 of the time. This means $R^2=0.9985$–0.9995, in which the variability has been compressed from 50% to 0.05%. For most workers it is harder to interpret the latter than the former. For example, addition of a parameter to a linear model will reduce s^2 when the new parameter is determined with standard error less than its magnitude [13]. Aware of this relation, a worker observing a modest decrease in s^2 on addition of, say, a quartic term to a quadratic model might choose the smaller-p model for its better extrapolating ability [20], but would opt for the expanded model on observing more significant decreases in s^2 [say by the factor $\nu/(\nu+3)$ when the new parameter is significant by 2σ]. Without translation to s^2, it is hard to distinguish between "modest" and "significant" increases in R^2.

Suppose the analyst applies the methods to a second analyte having sensitivity reduced by a factor of 3 (meaning slope b reduced by this factor). Comparable performance of the instrumentation and techniques will yield comparable s^2 but will now give a possibly alarming drop in R^2 to only "two nines", thanks to the order-of-magnitude drop in S_{YY} (which scales with b^2). Again, s^2 is the more intuitively meaningful metric.

Next consider weighted fitting. Assume we know the data σ_i absolutely and take $w_i=\sigma_i^{-2}$. Then for a correct fit model and normal error, s^2 in Eq. (8) becomes the reduced χ^2, $\chi^2_\nu=\chi^2/\nu$, which has expectation value 1. Thus we see that the "ideal" value of R^2 is not 1 but $1-1/S_{YY}$. The analyst in quest of maximal R^2 might fail to recognize the too-small values of χ^2 — probably from overestimates of the σ_i but perhaps from some other problem with the data. With χ^2 one also has a quick answer to the question, "Is the test model statistically reasonable?" — from χ^2 tables. Since every data set/fit model combination has its own S_{YY}, no such information is available for R^2 without first translating to χ^2. For weighted fits where the absolute scale is unknown, the χ^2 test cannot be used, and the situation is as for unweighted fitting.

4. Further comments

In illustrating the problems with R^2 under data transformation, Scott and Wild [4] produced two LLS fit models via different transformations,

$$Y_k = \alpha + \beta x + \gamma x^2, \quad (k = 1,2) \tag{9}$$

with $Y_1=y/x^{3/2}$ and $Y_2=\ln(y)$. Their use of unweighted LS was tantamount to different assumed data weightings, compounding the problems that exist even for proper weighting. A test of the two models under correct common weighting would have yielded valid R^2 comparisons and, of course, valid s^2 comparisons. For example, fitting to

$$y = \left(a + bx + cx^2\right)x^{3/2} \tag{10a}$$

and

$$y = \exp\left(A + Bx + Cx^2\right) \tag{10b}$$

would have accomplished this, with either $w_i=x_i^{-3}$ (as assumed for Y_1) or $w_i=y_i^{-2}$ (Y_2). However, neither R^2 nor s^2 (χ^2) can be used to decide which of these weighting choices is better, as the weights must be ascertained independently, preferably from replicate or pseudoreplicate measurements [21].

Scott and Wild did obtain nearly identical values for R^2 from Eq. (2) when they used their respective fit parameters to calculate the needed $\hat{y}_i$ for comparison on a common scale. But that comparison was specious, given the different tacit weightings in their two fits. Similarly, in comparing the several versions of Eq. (3), Bolster and Hornberger [10] correctly converted all their results to the form of Eq. (3) before computing their efficiency (R^2) values. However, in the fitting they did not weight the linearized versions in accord with the transformations. With proper weighting, there would have been no significant differences among the tested forms — indeed no numerical difference at all if the Deming-based algorithm were used.

In summary, R^2 *can* be used correctly for model comparisons when data are fitted to different models in fixed form with fixed weights. However, the comparisons are essentially obfuscated comparisons of s^2 or χ^2_ν, so why not just use these? In the past, data transformations have been widely used by physical scientists to achieve linear response functions, and by statisticians to achieve constancy of variance (justifying unweighted LS for the transformed data). However such transformations also alter the inherent distribution of the data; and especially in the case of physical data, seem more likely to render normal data nonnormal than the reverse. This is because many modern instruments either count or record averages of multiple digitizations of an analog signal. Counting yields Poisson data that are nearly Gaussian for large count, while averaging yields nearly normal data through the central limit theorem. With the capabilities of modern nonlinear LS algorithms, there is arguably little need for data transformations anyway. In particular, the realization that properly weighted fits to different mathematically equivalent versions of a given fit relation can yield a *single set of values* for parameters, standard errors, and χ^2 should forever end the misdirected practice of picking the "best" form from among such transformed relations — using R^2 or any other metric. Such efforts can better be focused on determining data error structures, from which reliable data weights can be assigned.

Acknowledgment

This research was part of USDA-ARS National Program 206: Manure and By-product Utilization.

References

[1] T.O. Kvålseth, Cautionary note about R^2, Am. Stat. 39 (1985) 279–285.
[2] J.B. Willett, J.D. Singer, Another cautionary note about R^2: its use in weighted least squares regression analysis, Am. Stat. 42 (1988) 236–238.
[3] E.L. Korn, R. Simon, Explained residual variation, explained risk, and goodness of fit, Am. Stat. 45 (1991) 201–206.
[4] A. Scott, C. Wild, Transformations and R^2, Am. Stat. 45 (1991) 127–129.
[5] R. Anderson-Sprecher, Model comparisons and R^2, Am. Stat. 48 (1994) 113–117.
[6] Y. Huang, N.R. Draper, Transformations, regression geometry and R^2, Comput. Stat. Data Anal. 42 (2003) 647–664.
[7] W. Huber, On the use of the correlation coefficient for testing the linearity of calibration functions, Accredit. Qual. Assur. 9 (2004) 726.
[8] D.B. Hibbert, Further comments on the (miss-)use of r for testing the linearity of calibration functions, Accredit. Qual. Assur. 10 (2005) 300–301.
[9] S.L.R. Ellison, In defence of the correlation coefficient, Accredit. Qual. Assur. 11 (2006) 146–152.
[10] C.H. Bolster, G.M. Hornberger, On the use of linearized Langmuir equations, Soil Sci. Soc. Am. J. 71 (2007) 1796–1806.
[11] R.J. Ritchie, T. Prvan, A simulation study on designing experiments to measure the K_m of Michaelis-Menten kinetics curves, J. Theor. Biol. 178 (1996) 239–254.
[12] A.G. Asuero, A. Sayago, A. Gonzalez, The correlation coefficient: an overview, Crit. Rev. Anal. Chem. 36 (2006) 41–59.
[13] P.R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969.
[14] D. Ruppert, N. Cressie, R.J. Carroll, A transformation/weighting model for estimating Michaelis-Menten Parameters, Biometrics 45 (1989) 637–656.
[15] J. Tellinghuisen, C.H. Bolster, Weighting formulas for the least-squares analysis of binding phenomena data, J. Phys. Chem. B 113 (2009) 6151–6157.
[16] W.E. Deming, Statistical Adjustment of Data, Wiley, New York, 1938,1943, Dover, New York, 1964.
[17] H.I. Britt, R.H. Luecke, The estimation of parameters in nonlinear, implicit models, Technometrics 15 (1973) 233–247.
[18] J. Tellinghuisen, Least-squares analysis of data with uncertainty in x and y: a Monte Carlo methods comparison, Chemom. Intell. Lab. Syst. 103 (2010) 160–169.
[19] J. Tellinghuisen, C.H. Bolster, Least-squares analysis of phosphorus soil sorption data with weighting from variance function estimation: a statistical case for the Freundlich isotherm, Environ. Sci. Technol. 44 (2010) 5029–5034.
[20] Q.C. Zeng, E. Zhang, J. Tellinghuisen, Univariate calibration by reversed regression of heteroscedastic data: a case study, Analyst 133 (2008) 1649–1655.
[21] J. Tellinghuisen, Variance function estimation by replicate analysis and generalized least squares: a Monte Carlo comparison, Chemom. Intell. Lab. Syst. 99 (2009) 138–149.