UNCERTAINTY

Uncertainty of measurements made by calibrated equipment

José Brenes André,
Escuela de Física, University of Costa Rica

The very important problem of finding the best fit line to interpret the readings of any calibrated equipment was addressed in the October 1998 issue of the OIML Bulletin by Dr. Subasinghe, who proposed a way to improve the calculation of uncertainties in such cases. The author of this paper would like to put forward a different approach which eliminates the obstacles encountered since the outset by using the least normal squares, a method which naturally allows the inversion of the independent and dependent variables, thus making it possible to calculate uncertainties using the known equations without having to make any simplifying assumptions. Although the equations derived here are known, the author has devised a simple way of his own to obtain them, which he believes makes the statistical properties associated with such a fit more evident. A practical application using the same GUM data as presented in the October 1998 Bulletin is also presented and discussed.

Introduction

In metrology, a very important case of a linear fit appears in the equipment calibration process, by using a given standard. A plot of the readings made by the equipment versus that obtained using the standard should give a straight line fit if the instrument is working properly, as Dr. Subasinghe points out in his paper. [1]

Although the author shares Dr. Subasinghe's main view that in the case of a plot of the equipment reading versus the reference equipment reading what is important is to find the uncertainty associated with a predicted x-value corresponding to a given y-value, he believes that rather than proposing a new way to approximate such uncertainty (equation 6 of his paper) the solution lies in finding a way to obtain the slope and intercept of the fit in such a way that, since the start, they fulfill the requirement of inversion.

In other words, rather than using the standard method known as Ordinary Least Square (OLS), and its result $m_{xy} = r^2 / m_{yx}$ (where r = the Pearson correlation coefficient), the Least Normal Squares (LNS) method should be used from the outset since it naturally leads to inversion, giving as a result $m_{xy} = 1 / m_{yx}$. (Caution: this is by no means to be taken as the special case r = 1.) This approach allows the user to solve the inversion problem without having to resort to the standard techniques of calculating uncertainties.

Ordinary least square (OLS) fit

The standard technique of the ordinary least square fit can be found in any introductory text, but for the sake of being comprehensive and to allow the reader to compare methods, it is outlined here.

Let (x_i, y_i) be a set of N ordered pairs, obtained experimentally. The values on the x-axis are supposed to have a negligible uncertainty. The best fit line will be characterized by that (m,b) which makes

$$S(m,b) = \Sigma (y_i - mx_i - b)^2$$

a minimum, which leads to:

$$m_{xy} = \frac{\sum (x_i - X)(y_i - Y)}{\sum (x_i - X)^2} \qquad b_{xy} = Y - m_{xy}\, X$$

where y_i and x_i denote the readings on the equipment needing corrections and those of the reference equipment, respectively. **X** and **Y** are the centroid of such a set of data points.

The suffix xy denotes that the variable x is considered to have been measured with negligible error, whereas the y variable not. For the case when all the errors have been lumped in the x variable the labels are inverted, and the following equations result:

$$m_{yx} = \frac{\sum (x_i - X)(y_i - Y)}{\sum (y_i - Y)^2}$$

$$b_{yx} = Y - m_{yx}\, X$$

$$m_{xy}\, m_{yx} = r^2$$

Least normal squares (LNS)

The different approach the author would like to present here is derived from a single tenet: the requirement that the dispersion $\Sigma\, v'^2_i$ (defined in the text below) be a minimum.

We start by shifting all the data values y_i and x_i by their respective centroid values, defining two new variables (u,v) such that $u_i = x_i - \mathbf{X}$, $v_i = y_i - \mathbf{Y}$. It is straightforward to show that $\mathbf{U} = 0$, $\mathbf{V} = 0$. Regardless of how the slope and intercept are calculated the best fit line has to pass through the centroid.

If the data shows a visible tendency to lie along a line, most of the data points can be encompassed by an imaginary ellipse, which suggests that the coordinates (u,v) should be rotated around the point (X,Y) by an angle α in a direction such that the variable u-axis will be practically parallel to the major semi-axis of the ellipse, which leads to:

$$u' = v \sin\alpha + u \cos\alpha \qquad v' = v \cos\alpha - u \sin\alpha$$

The value of α to be used will be that which makes the ellipse minor semi-axis a minimum. In the limiting case where all the points lie exactly on a line, the minor semi-axis will be exactly zero.

It is only natural to equate the dispersion of the u´ points to the value of the major semi-axis. The fact that the minor semi-axis is perpendicular to this axis is another way of saying that we are considering the distance from each point to the fit line as the value of a line segment starting from the point and ending in the fit line, when such a segment is perpendicular to the fit line.

Analysis in the (U´,V´) plane

The requirement that $\Sigma\, v'^2 = \Sigma(v\cos\alpha - u\sin\alpha)^2 = \Sigma\cos^2\alpha\,(v - u\tan\alpha)^2$ be a minimum is the equivalent of making S(m,b) a minimum with respect to m. The intercept b is in this case equal to zero, as can be seen from the proposed $v = u\tan\alpha$ relation.

In Anderson's paper [2], the angle between the line and the u-axis is taken as a parameter, but for other reasons (namely that the angle parameter transforms linearly whereas the m certainly does not). In the discussion that followed this article, P. Sprent states: "… there is a third idea that is unusual: that of considering the angle of inclination of a line to one of the axes rather than the tangent of that angle. I am not clear whether there are many practical advantages in looking at the angle rather than the slope" ([2], p. 20). This is an opinion the author of the present paper does not share.

It is apparent that:

$$\frac{du'}{d\alpha} = v' , \qquad \frac{dv'}{d\alpha} = -u',$$

from which the requirement of the dispersion $\Sigma\, v'^2$ being a minimum transforms to:

$$\frac{d}{d\alpha}\sum v_i^2 = -2\sum v_i\, u_i = 0$$

In other words, creation of the (u´, v´) plane is equivalent to making both new variables independent of each other, i.e. have zero correlation.

In [2] P. Sprent minimized $U = \Sigma\,(x\cos\varphi + y\sin\varphi - p)^2$ obtaining $\Sigma[(y - \mathbf{Y}) - \beta(x - \mathbf{X})][x + \beta y] = 0$. He drew attention to the fact that in Williams [3] two new variables $u = [(y - \mathbf{Y}) - \beta(x - \mathbf{X})]$ $v = [x + \beta y]$ are defined leading to $\Sigma\, uv = 0$ and so "the estimation of the slope β is equivalent to choosing β so that the sample correlation of (u,v) is zero" ([3] p.21).

Taking the derivative again, and forcing the dispersion of v´ to be a minimum one finds $-(\Sigma\, v'^2 - \Sigma\, u'^2) > 0$ which is another way of saying a > b, as it should be with the major and minor semi-axis. If instead one minimizes the dispersion of u´ , one finds that the assignment of signs is reversed, obtaining b > a opposite to our convention.

Analysis of (U,V) plane

Hence, for coherence we are forced to assign the major semi-axis to the dispersion of the u´ and the minor semi-axis to the dispersion of v´, from which:

$$a^2 = \frac{\sum u_i^2}{N} = \frac{\sum(u\cos\alpha + v\sin\alpha)^2}{N} = \frac{\sigma_y^2 + \sigma_x^2}{2} - \frac{\Delta}{2}$$

$$b^2 = \frac{\sum v_i^2}{N} = \frac{\sum(v\cos\alpha - u\sin\alpha)^2}{N} = \frac{\sigma_y^2 + \sigma_x^2}{2} + \frac{\Delta}{2}$$

The expressions using Δ are identical to equations (7) in Creasy [4].

We will denominate by φ the value of α, that minimizes b, or maximizes a, since both requirements are fulfilled simultaneously for the same angle, constituting further evidence of the coherence of this approach. This explains quite simply the result that the likelihood of the surface showing a saddle point rather than the absolute minimum as found out by Solari [6].

The above leads to:

$$\tan 2\varphi = \frac{2\sum vu}{\sum\left(u^2 - v^2\right)}$$

One can use this last equation to calculate the value of the slope tan φ by first finding φ and taking its tangent afterwards. However care has to be taken not to forget that there are two different possible values of φ (in the range $-\pi < \varphi < \pi$) for each value of tan 2φ. Such behavior is more clearly seen if the value of tan φ is directly obtained.

Using the facts that $\tan 2\alpha = 2\tan\alpha / (1 - \tan^2\alpha)$ and that we constructed the best fit line to be roughly parallel to the ellipse major semi-axis, we are allowed to identify tanα with the slope m, and hence find:

$$m_{xy} = \frac{B_{xy}}{r} \pm \sqrt{\left(\frac{B_{xy}}{r}\right)^2 + 1}$$

where we have set $B_{xy} = \frac{1}{2}\{(\sigma_y/\sigma_x) - (\sigma_x/\sigma_y)\}$. Notation reminds us that the first index for m refers to the independent variable, and the second to the dependent one.

Madansky [6] wrote: "One should note, though, that some of the papers referred to in Lindley [7] and Zucker [8] derive the least squares estimate of tan 2θ, where β = tanθ, when λ = 1, ..., but do not solve for β using the relation: $\tan 2\alpha = 2\tan\alpha / (1 - \tan^2\alpha)$".

And he continues: "Pearson [9] was one who estimated tan 2θ but he argued that the best-fitting straight line for the system of points coincides in direction with the major axis of the correlation ellipse. But the direction of the major axis of the correlation ellipse depends only on sgn (Σ xy). In none of the other papers do I find such an argument."

There are two possible values for m_{xy} depending on the sign used, which we will call m_{xy+} and m_{xy-}. Multiplication of both values gives -1 as a result, showing that they refer to two mutually perpendicular lines. Our convention forces us to assign the + sign to that parallel to the major semi-axis, leaving the – sign to that parallel to the minor semi-axis. Madansky [6] presents a very concise analysis of the use of the two signs.

Should we perform this analysis setting $\Sigma u'^2$ to be a minimum, then:

$$\sum u_1^2 = \sum(u\cos\alpha + v\sin\alpha)^2 = \sum(u + v\tan\alpha)^2 \cos^2\alpha$$

This implies that tanα is now the negative of the inverse of the previous tanα, which in turn corresponds to the second root.

We can also start from $a^2 > b^2$, $\Sigma u'^2 > \Sigma v'^2$ finding:

$$\sum u\,v \frac{\cos^2 2\alpha}{\sin 2\alpha} \succ -\sum u\,v \sin 2\alpha$$

Hence, if Σ uv > 0 we have to choose sin 2α > 0 which is fulfilled for $0 < \alpha < \frac{1}{2}\pi$, which is what we obtain if (x,y) are directly related. Similarly, if Σ uv < 0 we have to choose sin 2α < 0 which is fulfilled for $\frac{1}{2}\pi < \alpha < \pi$, which is what we obtain if (x,y) are inversely related. These results are coherent with the fact that with OLS method one finds $\tan\alpha = \Sigma vu / \Sigma u^2$ making it evident that the sign of Σ vu determines whether the data is directly or else inversely related.

The expressions for a^2 and b^2 represent a circle off-centered by $\frac{1}{2}(\sigma_y^2 + \sigma_x^2)$ in each axis, with a radius equal to 2R as given by equation (4) of Creasy [4], who uses this result to examine the confidence limits for the slope. The reader is referred to the original paperfor details.

Relation to OLS

Let us write:

$$\tan 2\alpha = \frac{2\sum u\,v}{\sum\left(u^2 - v^2\right)} = \frac{2\left(\sum uv \Big/ \sum u^2\right)}{1 - \dfrac{\sum v^2}{\sum u^2}}$$

Using $\tan 2\alpha = 2\tan\alpha / (1 - \tan^2\alpha)$ as a guide, one may be tempted to make the association $\tan\alpha = \Sigma vu / \Sigma u^2$, the result obtained when using the OLS method. But then one also has to take:

$$1 - \tan^2\alpha = 1 - \frac{\sum v^2}{\sum u^2} \quad \text{which leads to} \quad r^2 = 1$$

Such is the basis of Kendall's [10] equations (29.21) obtained by imposing certain conditions on the estimators. But from the above it can be seen that such an association can not be made, because it does not solve the quadratic equation. In fact equation (29.22) is basically $r^2 = 1$ which does not necessarily follow on from the first association.

Applying the approximation $\tan\beta \cong \beta$ when β is small one finds that $\tan 2\varphi \cong 2(\Sigma vu / \Sigma u^2)$ which implies that if the slope is small, the value obtained from the OLS and that obtained from the LNS will be basically the same (as expected) because in the OLS method one uses the differences in the y values which will now be practically perpendicular to the would be best fit line, as required by the LNS.

Calculation of the intercept

We obtained the value of the slope, but the intercept is still to be found. In the OLS case this is easily found from $\Sigma (y_i - m\, x_i - b) = \mathbf{Y} - m\, \mathbf{X} - b = 0$ which proves the assertion that the best fit line passes through the centroid.

For the LNS case we can use the equation for the slope to also find the value of b. Let us suppose that a new measurement (x_{N+1}, y_{N+1}) is made. The new average will equal the old one:

$$X_{new} = \frac{\sum x + x_{N+1}}{N+1} = \frac{N X + x_{N+1}}{N+1} = X$$

if and only if $x_{N+1} = \mathbf{X}$ and similarly for variable y. In the (u,v) plane the new data point is (0,0), i.e. the centroid. Because of the form of the tan 2φ equation, its value does not change. Hence, the centroid has to lie on the best fit line, and $\mathbf{Y} = m\, \mathbf{X} - b$ is also valid in the LNS approach.

Inversion

Basic to our claim is that the results can naturally be inverted, i.e. if we start with $\mathbf{Y} = m_{xy}\, \mathbf{X} + b_{xy}$ or else we start with $\mathbf{X} = m_{yx}\, \mathbf{Y} + b_{yx}$ we will always have $m_{xy} * m_{yx} = 1$. Exchanging x and y in the definition of the slope, it can be directly proved that $B_{xy} = -B_{yx}$, and that:

$$m_{xy}\, m_{yx} = \left[\frac{B_{xy}}{r} \pm \sqrt{\left(\frac{B_{xy}}{r}\right)^2 + 1}\right]\left[\frac{B_{yx}}{r} \pm \sqrt{\left(\frac{B_{yx}}{r}\right)^2 + 1}\right] = 1$$

The sign in front of the square root sign has to be the same in both cases, because all we did was to exchange x and y. Care has to be exercised not to confuse m_{yx} with the second root obtained from tan 2φ.

This result can also be obtained if one realizes that for a **y vs x** plot the slope m_{xy} is tanφ, whereas for an **x vs y** plot the slope m_{yx} will be tan $(\pi/2 - \varphi) = \cot\varphi$

Comparison of both methods

At this point a question arises: why switch from OLS to the alternative method presented here?

Using the same data presented in [1] for Observed temperature vs Reference temperature (Graph 2b: Calibration plot) (See Table 1), Dr. Subasinghe reports a linear relationship of:

$$y = 0.9978x + 0.2145$$

whereas using the method described here it is found that:

$$y = 1.00245\, x + 0.104$$

Although in this case the difference in the values of the slope is small, application of this method to other sets of data shows a more noticeable contrast. Moreover, from the epistemological point of view a method that makes no difference between which of the two variables is the dependent one and gives a single value for the slope is more satisfactory that one that forces the experimenter to make such a decision from personal considerations. ■

References

[1] Subasinghe, G.K.N.S. Expression of uncertainty in measurements made by calibrated equipment, OIML Bulletin No. 39 (1998) 18–23

[2] Anderson, T.W. (1976) Estimation of linear functional relationships: Approximate distributions and connections with simultaneous equations in Econometrics, J. R. Stats. Soc 38, 1–36.

[3] Williams, E. (1973) Tests of correlation in multivariate analysis. Bull. Int. Statist. Inst., Proceedings of the 39th Session, Book 4, 218–234

[4] Creasy, M.A. (1955) Confidence limits for the gradient in the linear functional relationship, J. Roy. Stats. Soc., 18, 65–69

[5] Solari, M. The maximum likelihood solution of the problem of estimating a linear functional relationship, J. R. Statist. Soc. B, 31 (1969), 372–375

[6] Madansky, A. The fitting of straight lines when both variables are subject to error, Am. Statis. Ass. J. 54 (1959), 173–205

[7] Lindley, D.V. Regression lines and the linear functional relationship, Jour. Royal stat. Soc, Supp, 9, (1947), 219–244

[8] Zucker, L.M., Evaluation of slope and intercept of straight lines, Human Biology, 19 (1947), 231–259

[9] Pearson, K. On lines and planes of closest fit to systems of points in space, Phil. Mag, 2 (1901), 559–572

[10] Kendall, M.G. and Stuart, A. (1977) The advanced theory of statistics, 2, 4th edition. London: Griffin

The author welcomes feedback to this article and may be contacted by e-mail: jbrenes@cariari.ucr.ac.cr