SPORTSCIENCE sportsci.org
Perspectives / Research Resources

Commentary on *Bias in Bland-Altman but not Regression Validity Analyses*

Alan M Batterham

Sportscience 8, 47-49, 2004 (sportsci.org/jour/04/amb.htm)
Department of Sport and Exercise Science, University of Bath, Bath BA2 7AY, UK. Email.
Reprint pdf · Reprint doc

The article and associated spreadsheet present a compelling argument for the use of ordinary least-squares regression analysis in calibration and validity studies. This communication was prompted by a spirited debate at the 2004 ACSM Annual Meeting and subsequently via electronic mail. It is worth stating that the analysis and interpretation of the study scenarios presented is one of the most controversial and often impenetrable topics in the measurement and biostatistics literature. The article discusses the most common scenario of two methods of measurement of some underlying quantity, with no replicate measurements. Several authors, including Bland and Altman (1999) and Dunn (2004) have argued that this design is weak as the resulting data are difficult to interpret with any confidence. These authors maintain that repeated measurements of each of the two methods should ideally be incorporated into the analysis, to take account for the influence of varying degrees of measurement error in both methods. This issue has implications for the potential artifactual proportional bias in Bland-Altman plots.

In the colloquium with Greg Atkinson that Will Hopkins refers to, we argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase. This, we suggested, could be due to the fact that one of the key assumptions of least-squares had been violated. In every statistical text and related article I have ever come across it is stated that the fitting of a straight line through least-squares assumes that there is only error in the Y variable (see, for example, Draper and Smith, 1998, for a detailed treatment). The X variable is assumed to be fixed by the experimenter or measured without error. This is because the squared deviations that are minimised are the residuals in the vertical (Y) direction only. Fitting by least-squares in situations where there is measurement error in X as well as Y reduces the slope of the line and also causes the intercept to deviate from zero. But as Will's analysis shows, using this line to calibrate X produces unbiased estimates of Y. Errors in X and Y contribute to the SEE from the least-squares model, of course, and the bigger the SEE the bigger the bias in a subsequent Bland-Altman plot. This can be demonstrated by altering the numbers in blue in the associated spreadsheet to reflect the preceding discussion. For instance, making the range large relative to the errors in X and Y removes the bias in the Bland-Altman plot. I used a mean of 100 and an SD of 30 for the true values, and errors of 5 and 1 for Y and X variables, respectively. When, however, the X or Y error is large relative to the range, the bias appears in the Bland-Altman plot.

At the ACSM colloquium, Greg Atkinson and I suggested that the bias in the Bland-Altman plot is not apparent if the original calibration was conducted via least-products regression. However, I acknowledge that there are other problems with this technique. As Will Hopkins argues, in the situations in which there is a known criterion or gold-standard, it seems inappropriate to "average" the two methods in this way. Also, least-products regression can lead to inflated SEEs and estimates that do not tend to their true values an N approaches infinity (Draper and Smith, 1998). Least-products may be useful in situations in which two methods are being compared and neither may be regarded as the gold standard. The debate then surrounds the prevalence of one measurement scenario (established criterion) versus the other (no gold standard measurement). John Ludbrook

and to a lesser extent Bland and Altman maintain that the latter is more common, whereas Will Hopkins argues that there is usually a criterion–otherwise, why would researchers conduct a validity study? Previously, I have expressed the view that much of the debate about appropriate analysis strategies is confused by various researchers' failing to distinguish between calibration/conversion problems, method comparison problems, and gold-standard method-comparison problems. It seems clear that limits of agreement and least-products regression may be suited to method comparison problems where there is no assumed gold-standard. In calibration/conversion problems where "Y" is the criterion and "X" is the cheap, practical alternative, ordinary regression techniques may suffice, as Will Hopkins suggests. However, if these are conducted using least-squares, and a subsequent validity analysis is conducted via Bland-Altman plots, an apparent proportional bias is inevitable unless the X-range is very large relative to errors in X and Y (or small error in Y, and zero X-error – as per least-squares assumptions). My take, currently, is that least-squares regression may be used at the validity stage, as Will Hopkins suggests, especially when the measurement scenario involves a gold-standard. I do not believe that Bland and Altman would disagree with this position. Indeed, as pointed out and illustrated at the aforementioned ACSM colloquium, these authors have recently proposed that regression may be used on the original, raw variables to provide something akin to the limits of agreement (Bland and Altman, 2003). At the colloquium, Greg and I proposed 90% limits for the prediction error from a linear regression for an individual at the mean of the distribution (approximately 1.65*SEE). As Will Hopkins has suggested previously, however, 68% limits for the SEE may be more meaningful for clinical or practical significance.

Will Hopkins has contributed an illuminating addition to the ongoing analysis debate with section and spreadsheet update on the application of Bland-Altman plots to measures that have not been calibrated through least-squares, and that differ only in random error. The suggestion is that the bias is still apparent, and by implication the bias is, therefore, is not an artefact of least-squares at the calibration stage as no calibration occurred. However, if one experiments with the range of the true values as defined by the mean and SD, and makes the errors in each of the two methods small relative to the range (even with one large relative to the other), it seems that the bias in the Bland-Altman plot reduces substantially or disappears, as in my previous example. This suggests that this bias, too, is related to the magnitude of the errors relative to the range of the data.

In summary, I believe that with this article and spreadsheet Will Hopkins has made an excellent contribution to the literature that should stimulate further debate. I believe that slavish adoption to any measurement technique, including limits or agreement, should be discouraged. I would urge researchers where possible to conduct and incorporate a thorough analysis of measurement error into their studies, and to identify the measurement scenario as method comparison, comparison against a gold standard, and/ or calibration/ conversion. As Will Hopkins has proposed, recalibration of instruments using least-squares regression at the validity phase can be useful. Dunn (2004) echoes that Y values can be re-scaled at the validity stage prior to additional analysis. I am not quite ready to state that Bland-Altman plots and limits of agreement are totally redundant, but I agree that standard regression techniques can be employed to analyze calibration, conversion, and validity studies appropriately, especially in situations in which there is a known or assumed criterion or gold-standard, and when measurement error in X and Y is small relative to the range of the data.

References

Bland JM, Altman DG (1999). Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8, 135-160

Bland JM, Altman DG (2003). Applying the right statistics. Ultrasound in Obstetrics and Gynecology 22, 85-93

Dunn G (2004). Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies, 2nd ed. London: Arnold

Draper NR, Smith H (1998). Applied Regression Analysis, 3rd ed. New York: Wiley

Ludbrook J (1997). Comparing methods of measurement. Clinical and Experimental Pharmacology and Physiology 24, 193-203

Back to article/homepage

Published Nov 2004