A linear regression approach to check bias between methods – Part II

A Worked Example

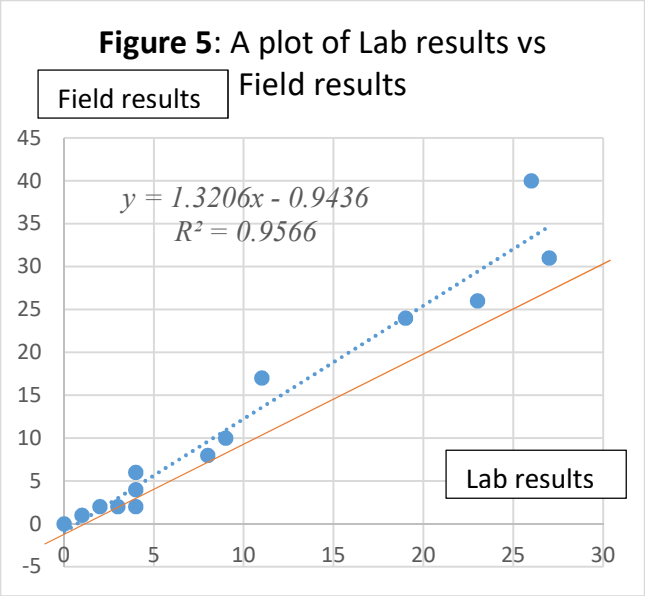
Suppose that we determined the amount of uranium contents in 14 stream water samples by a well-established laboratory method and a newly-developed hand-held rapid field method [1]. The purpose is to test the accuracy of the field method, regarding the laboratory method as a reference point of accuracy, being more precise than the field method does. The results, in units of μg/L are as follows:

Table 1: Uranium analysis of 14 stream

Water samples in units of μg/L

	x	y
Site code	Lab result	Field result
1	19	24
2	8	8
3	2	2
4	1	1
5	9	10
6	23	26
7	27	31
8	11	17
9	4	4
10	0	0
11	4	6
12	26	40
13	3	2
14	4	2

Figure 5: A plot of Lab results vs Field results



Legends:

..... the regression line

——— the theoretical line of no bias

The following equations were used to calculate various statistical parameters for a linear regression equation $y = a + bx$ before carrying out the Student's t significance testing:

1. The slope (gradient), $b = 1.321$ which agreed well with the linear equation produced by the MS Excel software.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2. The y–intercept, $a = -0.944$ which also agreed with the MS Excel calculation

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

3. The correlation coefficient, $r = 0.978$ and $R^2 = 0.957$

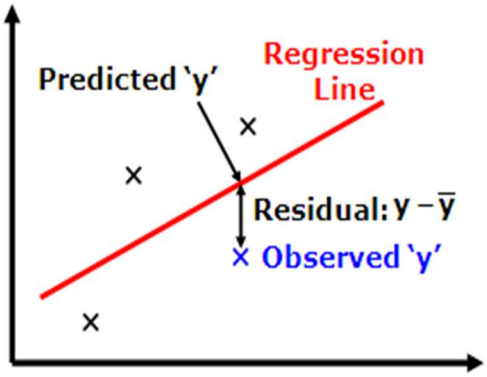
$$r = b\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - y)^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Hence, the linear regression equation for this set of data is $y = -0.944 + 1.321x$

In addition, we have to carry out a residual analysis on the data to obtain the standard error of y on x, the standard deviation of the slope, s_b, and the standard deviation of the y – intercept, s_a.

As mentioned earlier, a residual (which is also called regression error) at point x_i is defined as: $y_i - \hat{y}_i$ where y_i is the experimental or observed value of y at that point x_i, whilst $\hat{y}_i$ is the fitted or predicted value of y at point x_i, based on the linear regression equation. See Figure 6 below.

Figure 6: A graphic picture of residual



Sum of squared errors (squared residuals) is $SSe = \sum(y_i - \hat{y}_i)^2$ and the degree of freedom, v is (n–2) where n is the number of sets of (x_i,y_i). The mean squared error therefore is:

$$MSe = \frac{SSe}{v} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

and, the standard error of y on x is:

$$s_{y/x} = \sqrt{MSe} = \sqrt{\frac{SSe}{v}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

The standard error of y on x, $s_{y/x}$ is a very important component in the estimation of linear regression, i.e. variance for the slope b and the intercept a, as shown in the following equations:

$$s_b = \frac{s_{y/x}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$s_a = s_{y/x}\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}} = s_b\sqrt{\frac{\sum x_i^2}{n}}$$

In this example, we have found:

$s_{y/x}$ = 2.816; s_b = 0.081; s_a = 1.111

Now, we shall carry out a hypothesis testing at probability α = 0.05 level as described below:

Ho: $b = 1$; $a = 0$

H1: $b \neq 1$; $a \neq 0$

The Student's t-test with the following general equation is used for such hypothesis or significance testing:

$$t = \frac{|\mu - \bar{x}|\sqrt{n}}{s}$$

For the value of t_b, we have: $t_b = \frac{|1-b|\sqrt{n}}{s_b}$ = 3.95 with p = 0.002 < 0.05

and, for the value of t_a, we have: $t_a = \frac{|0-a|\sqrt{n}}{s_a}$ = 0.85 with p = 0.411 > 0.05

The critical value for t at (14-1) or 13 degrees of freedom at 95% confidence level is 2.16.

In conclusion, the analysis of significance testing results shows that the intercept is not significant from zero (p = 0.412 > 0.05) but the slope of 1.32 is clearly significantly different from unity (1), an important requirement for no bias of any kind between the methods. The field method in this case is giving results which are on average 1.32 times greater than the established laboratory method.

Estimation of bias uncertainty

Last but not least, we will estimate the uncertainty of relative bias, $u_{b,rel}$ which has to be added as another component in the overall evaluation of measurement uncertainty.

In this example, we have paired results from two separate methods. We first calculate their individual difference, $d_i = x_i - y_i$ and then its relative difference in relation to the corresponding lab result, which was taken as the reference, i.e. $d_{i,rel} = \frac{d_i}{x_i} = \frac{x_i - y_i}{x_i}$

The relative standard uncertainty of bias is given by the equation:

$$u_{b,rel} = \sqrt{\frac{\sum d_{i,rel}^2}{n}}$$

The calculated differences, d, are tabulated as below:

	x	y			
Site code	Lab result	Field result	Difference d_i	$d_{i,rel}$	$d_{i,rel}^2$
1	19	24	-5	-0.263	0.069
2	8	8	0	0.000	0.000
3	2	2	0	0.000	0.000
4	1	1	0	0.000	0.000
5	9	10	-1	-0.111	0.012
6	23	26	-3	-0.130	0.017
7	27	31	-4	-0.148	0.022
8	11	17	-6	-0.545	0.298
9	4	4	0	0.000	0.000
10	0	0	0	0.000	0.000
11	4	6	-2	-0.500	0.250
12	26	40	-14	-0.538	0.290
13	3	2	1	0.333	0.111
14	4	2	2	0.500	0.250

Hence, $u_{b,rel} = \sqrt{\frac{\sum d_{i,rel}^2}{n}} = \sqrt{\frac{1.319}{14}} = 0.307$

References:

[1] Experimental data taken from Michael Thompson & Philip Lowthian, *Notes on Statistics and Data Quality for Analytical Chemists*, Imperial College Press, 2011