XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

DIFFERENT METHODS OF LINEAR FIT IN $\mathbb{R}^2$

Wilhelm Kolaczia

Section of Dimensional Quantities, Photometry
BEV - Bundesamt für Eich- und Vermessungswesen
Arltgasse 35, A-1160 Wien, Austria
E-mail: w.kolaczia@metrology.at

Abstract – In the field of metrology it is common practice to approximate the results of a measurement by functions modelling physical facts. The most simple form of such an approximation is the linear fit. This article presents five different methods of linear fit in $\mathbb{R}^2$ by working out the characteristic features of these methods like scale unit invariance. The author´s intention is to be helpful in finding the most useful method from case to case.

Keywords: Linear fit, Scaling, Scale unit invariance

1. Linear fit with respect to *x* or to *y*

The method of linear fit in $\mathbb{R}^2$ used in most cases is to find a function $g(x)$ for a given set $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ of ordered pairs for which

$$\sum_{i=1}^{n}\left(y_i - g(x_i)\right)^2 \quad \text{with } g(x_i) = kx_i + d \tag{1}$$

is a minimum. This condition is fulfilled for

$$k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad d = \bar{y} - k\bar{x} \tag{2}$$

whereas $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ applies [2].

The ordered pairs (x_i,y_i) can be interpreted geometrically as points and the function $g(x)$ as straight line in the plane, which can be described by the equation $y-\bar{y}=k(x-\bar{x})$. This line minimizes the vertical distances and will be called g_y with slope k_y.

Example:
$M = \{(2,4;1,15),(4,8;2,95),(7,2;6,4),(9,6;5,05),(12,0;6,95)\}$. The fitted straight line g_y (fig. 1) fulfils for the set M the equation

$$y - 4,50 = 0,571(x - 7,20).$$

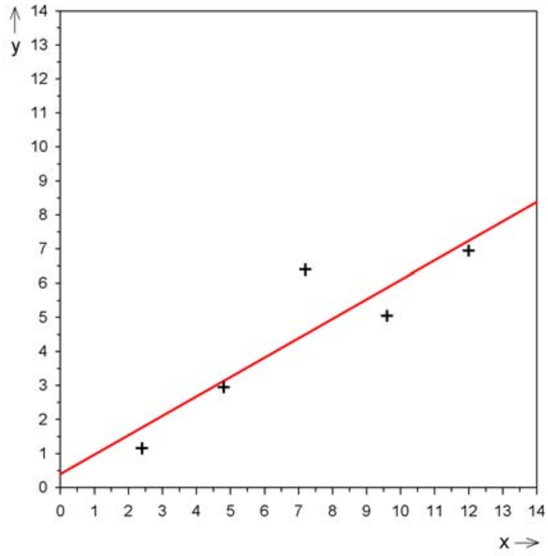

Figure 1: Fitted straight line g_y

Analogously, one can find a straight line g_x, which minimizes the horizontal distances. This line fits the equation $y - \bar{y} = k_x\,(x - \bar{x})$, whereas

$$k_x = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})} \tag{3}$$

applies [2].

The fitted straight line g_x (fig. 2) fulfils for the set M (chosen above as an example) the equation

$$y - 4,50 = 0,716\,(x - 7,20)$$

where $k_y \le k_x$ applies (fig. 3); in general [2], $\left|k_y\right| \le \left|k_x\right|$ and

$$\text{sign}\left(k_x\right) = \text{sign}\left(k_y\right) = \text{sign}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) \text{ applies.}$$

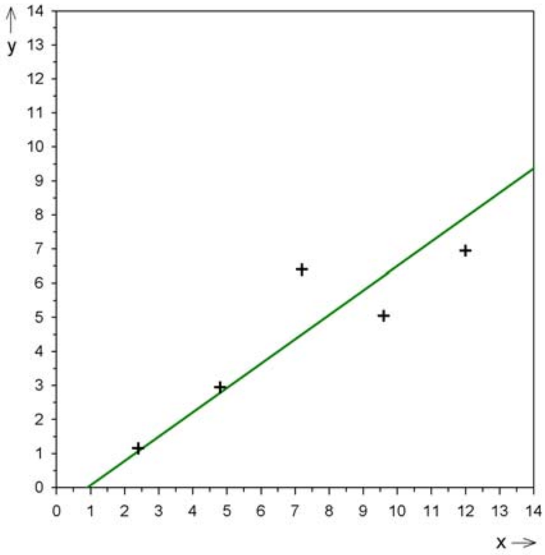

Figure 2: Fitted straight line g_x

In order to simplify these expressions a little bit more, the following notation shall be used:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2,$$
$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{4}$$

Therefore, k_x and k_y can be expressed as follows:

$$k_x = \frac{S_{yy}}{S_{xy}} \quad \text{and} \quad k_y = \frac{S_{xy}}{S_{xx}} \tag{5}$$

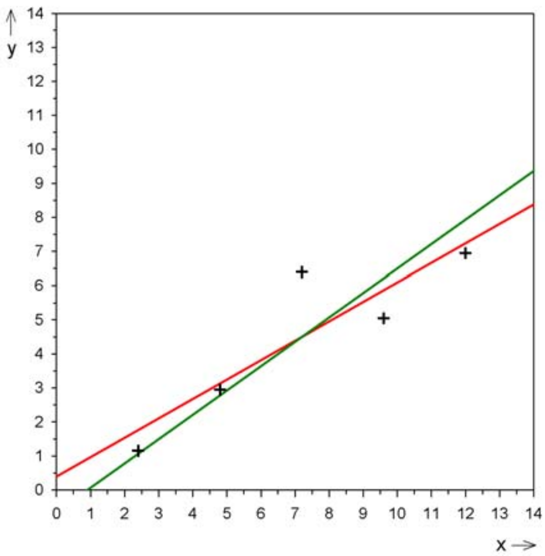

Figure 3: Fitted straight lines g_x and g_y

2. Linear fit with respect to *x* and to *y*

The fitted straight lines g_x and g_y fit only with respect to *x* or to *y*. Sometimes it is appropriate to fit with respect to *x* and also to *y*. The straight line g_{xy} shows this property minimizing the normal distances. It fits the equation $y - \overline{y} = k_{xy}(x - \overline{x})$, whereas

$$k_{xy} = \frac{-\left(S_{xx} - S_{yy}\right) + \sqrt{\left(S_{xx} - S_{yy}\right)^2 + \left(2S_{xy}\right)^2}}{2S_{xy}} \tag{6}$$

applies [1-2].

Furthermore, this fit with respect to *x* and also to *y* minimizes absolutely the sum of the squared deviations of the points given and the fitted straight line [1-3]. This method is called the *total least-square fit*.

The fitted straight line g_{xy} (fig. 4) fulfils for the set *M* (chosen above as an example) the equation

$$y - 4{,}50 = 0{,}608\,(x - 7{,}20).$$

where $k_y \le k_{xy} \le k_x$ applies (fig. 5); in general [2],

$$|k_y| \le |k_{xy}| \le |k_x| \quad \text{and}$$
$$\text{sign}(k_x) = \text{sign}(k_y) = \text{sign}(k_{xy}) = \text{sign}(S_{xy}) \quad \text{applies.}$$

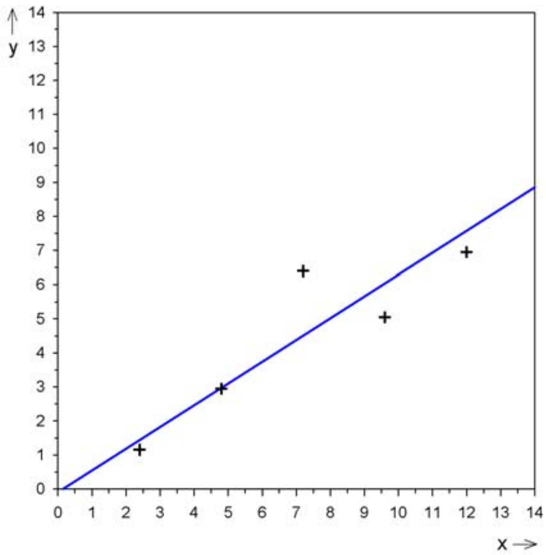

Figure 4: Fitted straight line g_{xy}

As it will be shown in the following both coordinates are usually not taken into account in the same way. As it can be seen in figure 6 the distance *Δ* of a point (x_i, y_i) from a given straight line g can be calculated as follows:

$$\Delta = \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{7}$$

whereas $\Delta x_i = x_i - x$ and $\Delta y_i = y_i - y$ with $(x,y) \in g$ applies.

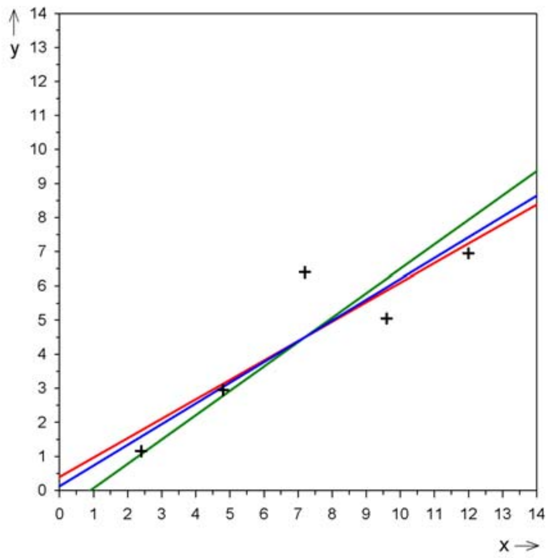

Figure 5: Fitted straight lines g_x, g_y and g_{xy}

If $\Delta x_i \neq \Delta y_i$ applies, both coordinates will influence the distance Δ in a different way, whereas the coordinate with the bigger coordinate distance is stronger taken into account.

If $|\Delta x_i| > |\Delta y_i|$ applies, then the *x*-coordinate will be stronger taken into account, if, to the contrary, $|\Delta x_i| < |\Delta y_i|$ applies, then the *y*-coordinate will be stronger taken into account. In the case of figure 6 this is the *x*-coordinate.

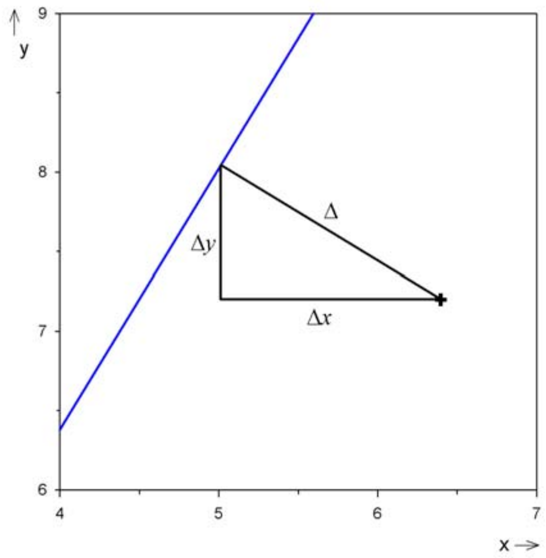

Figure 6: distance Δ of a point (x_i,y_i) from a given straight line *g*

For the slope *k* of the straight line applies:

$$k = -\frac{\Delta x}{\Delta y} \tag{8}$$

This leads for the fitted straight line g_{xy} to the following relations:

If $|k_{xy}| < 1$ applies, then g_{xy} fits stronger with respect to *y* than to *x*;

If $|k_{xy}| = 1$ applies, then g_{xy} fits equally with respect to *x* and to *y*;

If $|k_{xy}| > 1$ applies, then g_{xy} fits stronger with respect to *x* than to *y*;

3. Scaling properties of the fitted straight lines

If a linear fit made uniformly with respect to x and y is wanted the straight line g_{xy} is, generally speaking, not useful. Furthermore, the method of minimizing the normal distances has an extremely bad property which makes an application in the measurement techniques questionable: The result received depends on the choice of the unit of measurement in such a way as it is not possible to convert it later into any other unit. E.g., if distances are measured in metres and times in seconds in order to determine the velocity of a body in uniform motion and a linear fit by the method of minimizing the normal distances is carried out and the result shall be given in kilometres per hour, then this result will be different from that, which would have been received, if the fit would have been made on base of the numerical values already converted before in kilometres and hours, respectively.

The conversion from one unit to another one is – in the mathematical sense – a transformation of coordinates called *scaling* and is defined as follows:
Be $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ a set of ordered pairs, c_x and c_y real numbers and *S* an operator with the property

$$S(M) = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)\} = M^* \\ \text{with } x_i^* = c_x x_i \text{ and } y_i^* = c_y y_i, \tag{9}$$

then *S* will be called *scaling operator*, M^* *scaling of M under S*, both constants c_x and c_y *scaling coefficients of x and y* and their quotient $\frac{c_y}{c_x}$ the *scaling factor*.

If $\frac{c_y}{c_x} \neq 0$, applies, then an scaling operator S^{-1} exists with a scaling factor $\left(\frac{c_y}{c_x}\right)^{-1}$ with the property

$$S^{-1}(S(M)) = \left(S \circ S^{-1}\right)(M) = M \tag{10}$$

So, the transformation of coordinates by a scaling *S* is reversible.

Be $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ a set of ordered pairs and *g*(*M*) a fitted line in *M*, then *g*(*M*) is called *unit invariant with respect to a transformation of coordinates* by the scaling operator *S* if

$$g(S(M)) = S(g(M)) \text{ applies.} \tag{11}$$

The fitted straight lines g_x and g_y are scale unit invariant with respect to scaling under *S* because

$$k_{x^*} = \frac{c_y}{c_x} k_x \text{ and } k_{y^*} = \frac{c_y}{c_x} k_y \text{ applies [2].} \tag{12}$$

To the contrary, the fitted straight line g_{xy} is not scale unit invariant with respect to scaling under S because

$$k_{x^*y^*} = \frac{c_y}{c_x} k_{xy} \Leftrightarrow \left|\frac{c_y}{c_x}\right| = 1 \qquad (13)$$

applies [2].

So the total least-square fit method which follows optimaly the Maximum-Likelihood-principle [3] is not very useful in the field of metrology.

4. Linear fit with respect to *x* and to *y* using the slopes k_x and k_y

If one carries out a transformation of coordinates of the set $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ by scaling operator S with the scaling coefficients

$$c_y = \sqrt{S_{xx}} \text{ and } c_x = \sqrt{S_{yy}} \qquad (14)$$

and calculates afterwards the fitted line $g_{x^*y^*}$ with respect to x and to y for $S(M) = M^*$, so one gets the equation $y - \bar{y} = k_{x^*y^*}(x - \bar{x})$ for the fitted line $g_{x^*y^*}$ whereas

$$k_{x^*y^*} = \operatorname{sign}(S_{xy}) \text{ applies [2].} \qquad (15)$$

Calculating the fitted lines for M^* with respect to x and to y leads to [2]:

$$k_{y^*} = r_{xy} \text{ and } k_{x^*} = k_{y^*}^{-1} \qquad (16)$$

whereas r_{xy} is the correlation coefficient of x and y.

These relations are important since under the conditions given the fitted line $g_{x^*y^*}$ is the axis of symmetry of the angle between both fitted lines g_{x^*} and g_{y^*}. Furthermore, because of $\left|\operatorname{sign}(S_{xy})\right| = 1$ $g_{x^*y^*}$ fits uniformly in M^* with respect to x and to y.

Therefore, it is obvious – following the proposal of *Danzer* [4] – to apply the axis of symmetry of the angle between both fitted lines g_x and g_y as fitted straight line $\tilde{g}_{xy}$ with respect to x and to y. It fits the equation $y - \bar{y} = \tilde{k}_{xy}(x - \bar{x})$, whereas

$$\tilde{k}_{xy} = \tan\frac{\arctan k_x + \arctan k_y}{2} \qquad (17)$$

applies [4]. In any set of ordered pairs $\tilde{g}_{xy}$ fits uniformly with respect to x and to y.

For the set M (chosen above as an example) the fitted straight line $\tilde{g}_{xy}$ (fig. 7) fulfils the equation

$$y - 4{,}50 = 0{,}641\,(x - 7{,}20).$$

An argument against this kind of linear fit is that the fitted straight line as defined above is not invariant against a transformation of coordinates by a scaling S which shall be shown below.

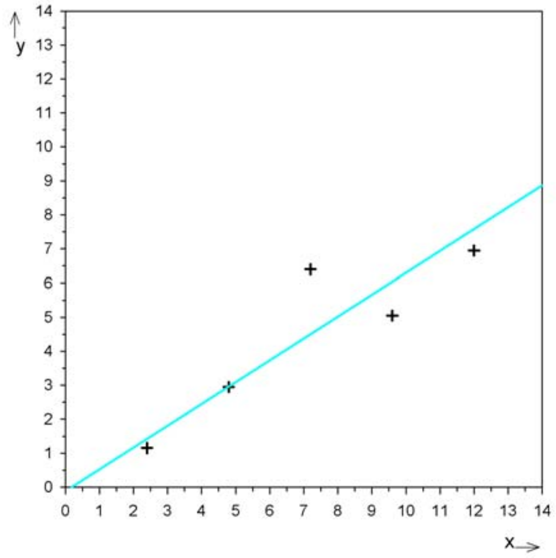

Figure 7: Fitted straight line $\tilde{g}_{xy}$

Be $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ a set of ordered pairs and S the scaling operator with the scaling coefficients given in (11). Carrying out a transformation of coordinates of M by scaling operator S and calculating afterwards the fitted line $g_{x^*y^*}$ with respect to x and to y for $S(M) = M^*$ gives equation (12). A further transformation of coordinates in $M^* \cup g_{x^*y^*}$ by the scaling operator S^{-1} with the scaling coefficients

$$c_{y^*} = \sqrt{S_{yy}} \text{ and } c_{x^*} = \sqrt{S_{xx}} \qquad (18)$$

leads to a straight line $\hat{g}_{xy} : y - \bar{y} = \hat{k}_{xy}(x - \bar{x})$ in M whereas

$$\hat{k}_{xy} = \operatorname{sign}\left(S_{xy}\right)\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} \qquad (19)$$

applies [2].

For the set M (chosen above as an example) the fitted straight line $\hat{g}_{xy}$ (fig. 8) fulfils the equation

$$y - 4{,}50 = 0{,}639\,(x - 7{,}20)$$

Analogously one also can find

$$\hat{k}_{xy} = \operatorname{sign}\left(S_{xy}\right)\sqrt{k_x k_y} \qquad (20)$$

while taking into account the relationship

$$\frac{S_{yy}}{S_{xx}} = \frac{S_{yy}}{S_{xy}} \cdot \frac{S_{xy}}{S_{xx}} = k_x k_y \qquad (21)$$

The slope $\hat{k}_{xy}$ of the straight line $\hat{g}_{xy}$ is as far as the amount is concerned the geometrical mean of the slopes k_x and k_y.

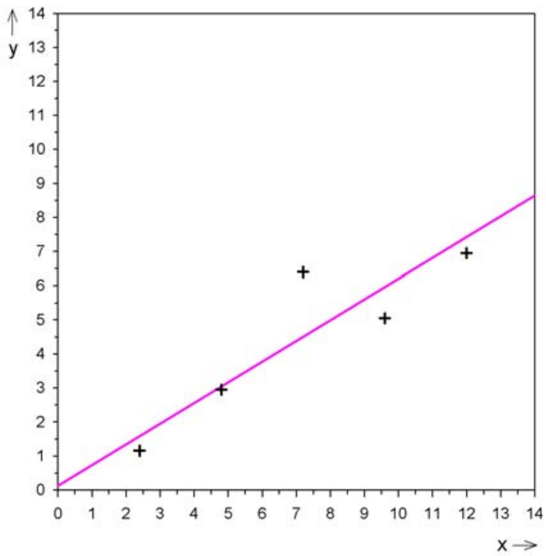

Figure 8: Fitted straight line $\hat{g}_{xy}$

The straight line $\hat{g}_{xy}$ fits in *M* with respect to *x* and to *y* and is unit invariant against a transformation of coordinates by scaling operator *S*, since

$$\hat{k}_{x^*y^*} = \frac{c_y}{c_x}\hat{k}_{xy} \tag{22}$$

applies [2].

As it was seen above both straight lines $\tilde{g}_{xy}$ and $\hat{g}_{xy}$ are only slightly different; generally speaking, they follow the relationship

$$\hat{g}_{xy} = \tilde{g}_{xy} \Leftrightarrow k_x = k_y \vee k_x = k_y^{-1} \tag{23}$$

Therefore, the straight line $\hat{g}_{xy}$ fits uniformly with respect to *x* and *y*, if

$$k_x = \hat{k}_{xy} = k_y \vee \hat{k}_{xy} = \mathrm{sign}(S_{xy}) \text{ applies.} \tag{24}$$

Be $M = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$ a set of ordered pairs, for which $k_x = k_y^{-1}$ applies. So, also $\hat{g}_{xy} = \tilde{g}_{xy}$ and consequently $\hat{k}_{xy} = \tilde{k}_{xy} = \mathrm{sign}(S_{xy})$ applies.
If a transformation of coordinates is carried out with *M* by a scaling *S* with the scaling coefficients c_x and c_y, whereas $|c_x| \neq |c_y|$ applies, the result is $k_{x^*} \neq k_{y^*}^{-1}$ and hence

$$\hat{k}_{x^*y^*} \neq \tilde{k}_{x^*y^*}. \tag{25}$$

On the other hand

$$\frac{c_y}{c_x}\tilde{k}_{xy} = \frac{c_y}{c_x}\hat{k}_{xy} = \hat{k}_{x^*y^*} \tag{26}$$

applies, and therefore altogether

$$\tilde{k}_{x^*y^*} \neq \frac{c_y}{c_x}\tilde{k}_{xy} \tag{27}$$

which demonstrates that the fitted straight line $\tilde{g}_{xy}$ is not invariant against a transformation of coordinates by a scaling *S*.

For the set *M* (chosen above as an example)

$$k_y \leq k_{xy} \leq \hat{k}_{xy} \leq \tilde{k}_{xy} \leq k_x \text{ applies.}$$

5. Conclusion

All the presented fitted straight lines fulfil the equation $y - \bar{y} = k(x - \bar{x})$, whereas

$$k_y \leq k \leq k_x \tag{28}$$

applies (fig. 9).

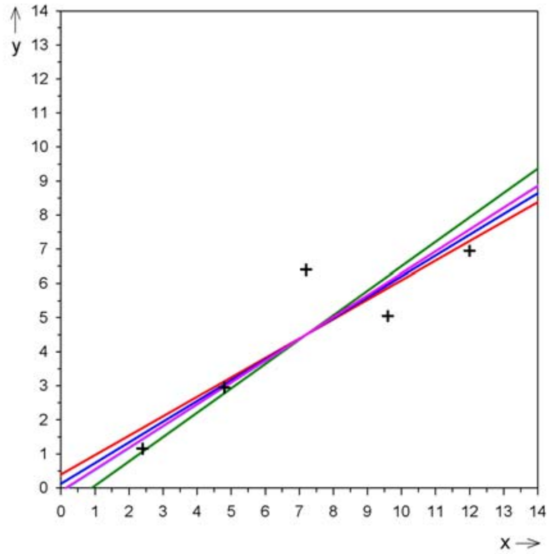

Figure 9: Fitted straight lines g_x, g_y, g_{xy}, and $\hat{g}_{xy} \approx \tilde{g}_{xy}$

The slopes $k_x, k_y, k_{xy}, \tilde{k}_{xy}$ and $\hat{k}_{xy}$ of the fitted lines $g_x, g_y, g_{xy}, \tilde{g}_{xy}$ and $\hat{g}_{xy}$ have the following relationship:

$$|k_y| \leq |k_{xy}| \leq |\hat{k}_{xy}| \leq |\tilde{k}_{xy}| \leq |k_x| \text{ for } |\tilde{k}_{xy}| < 1 \text{ and}$$

$$|k_y| \leq |\tilde{k}_{xy}| \leq |\hat{k}_{xy}| \leq |k_{xy}| \leq |k_x| \text{ for } |\tilde{k}_{xy}| > 1 \text{ respectively.}$$

The straight lines g_x and g_y fit one-sided – either with respect to *x* or to *y* – but are on the other hand invariant against a transformation of coordinates by a scaling *S*.

The straight line g_{xy} fits – although not uniformly – with respect to *x* and to *y*, but is not invariant against a transformation of coordinates by a scaling *S*.

The straight line $\tilde{g}_{xy}$ fits uniformly with respect to *x* and to *y*, but is also not invariant against a transformation of coordinates by a scaling *S*.

The straight line $\hat{g}_{xy}$ fits, although not as uniformly as the straight line $\tilde{g}_{xy}$ does, with respect to *x* and to *y* – but is invariant against a transformation of coordinates by a scaling *S*.

In the field of metrology the invariance against the choice of the unit of measurement is absolutely necessary. Therefore, the straight lines g_x, g_y and $\hat{g}_{xy}$ are very useful for a linear fit. If a linear fit with respect to *x* and to *y* is necessary – for any reason so far, the fitted line $\hat{g}_{xy}$ is the best possible compromise.

REFERENCES

[1] M. Krystek: Ausgleichsgeraden in der Ebene. In: tm – Technisches Messen 71 (2004) Nr. 1, S. 19-23.

[2] W. Kolaczia: Approximation und Interpolation: Lineare Regression und Korrelation. BEV, Wien 2005.

[3] A.Balsamo, G. Mana und F. Pennechie: On the best fit of a line to uncertain observation pairs. In: Metrologia **42** (2005) 376-382.

[4] K. Danzer, L.A. Currie: Guidelines for Calibration in Analytical Chemistry. Part 1. In: Pure & Appl. Chem. 70 (1998), Nr. 4, S. 993-1014.