Confidence intervals for weighted polynomial calibrations

Sergey Maltsev, Ampersand Ltd., Moscow, Russia;
Yuri Kalambet, Ampersand International, Inc., Beachwood, OH
e-mail: kalambet@ampersand-intl.com
http://www.chromandspec.com

Summary

In many chromatographic software products, weighted polynomial regression is used for the calibration curves. Weighting is a useful mean to account for the measurement error that may depend on the detector response value. Confidence intervals show how accurate a measurement is with the help of a calibration curve.
We have extended the confidence interval theory to the frequent case of weights, expressed as a function of Y value, in particular 1/Y and $1/Y^2$. This extension allows accurate calculation of concentrations with confidence intervals for calibration curves, constructed using point weighting. Examples are shown that demonstrate applications of weighted calibration curves with confidence intervals in chromatography.

Introduction

General theory of linear regression analysis is usually used for calculating confidence intervals. However, in analytical chemistry, we may face a practical problem where the regression analysis cannot be applied directly and needs some adaptation.

A very common situation in analysis is that the error is not the same for different signal levels. The error can be proportional to the signal itself, that is $\Delta_R \sim R$. For radioactivity measurements we would have $\Delta_R \sim \sqrt{R}$. Here, R is the detector response and Δ_R is a related measurement error. Correct handling of such errors and calculating true errors in the resulting concentration requires special consideration.

Another example of weighted regression is curve fitting by polynomial of second or third power. Conventional linear regression analysis gives tools for calculating confidence intervals for this case. Exact formulas are typically not present in specialized literature for analytical chemistry.

The most important practical cases related to the confidence intervals are the following:

1. We build the regression of value y_i which is measured with error at the precisely known set x_i (calibration). Then, we make a set of measurements of the detector response Y_* at known x_*. What is the variation of Y_* value and how does it differ from the expected value $\hat{Y}_*$, calculated from calibration at x_*?
2. We build the regression of value y_i which is measured with error over the precisely known set x_i (calibration). Then, we measure the detector response Y_* at some unknown x_*. A value $\hat{x}_*$ from calibration curve at Y_* is used as the estimate of x_*. What is a variation of $\hat{x}_*$ value and how it differs from "true" x_*?

This article is based on the classical work on linear regression analysis [1]. Still applied aspects of analytical chemistry are considered.

I. Regression Without Weighting

Regression without weighting assumes that the error of measurement depends on neither the detector response nor the concentration. Regression can be linear through origin or not through origin. Also, it can be a polynomial of second or third order.
This case is considered in details in the literature.

A regression is defined by the expression:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{(I.1), (Seber 3.2)}$$

where $\mathbf{X}$ is a regression matrix. For example, quadratic regression not going through origin it is

$$\mathbf{X} = \begin{Bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{Bmatrix}$$

The lower index refers to the number of calibration point.
$\mathbf{Y} = \{y_1.y_2...y_n\}$ is vector of detector response values
$\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$ - a related vector of errors of measurements.
$\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\}$ - a regression coefficients which must be calculated.

It is assumed that errors of measurements follow the rule:
$\mathrm{cov}[\varepsilon_i, \varepsilon_j] = \delta_{ij} \cdot \sigma^2$
$\mathbf{D}[\varepsilon] = \sigma^2 \cdot \mathbf{I_n}$
That is, errors ε_i are not correlated and have the same dispersion.
The solution of the dispersion is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \cdot \mathbf{X})^{-1} \mathbf{X}' \cdot \mathbf{Y} \qquad \text{(I.2), (Seber 3.5)}$$

Confidence interval at $100 \cdot (1-\alpha)$ **level for response value at a given** x_*

We are building the regression of response values y_i measured with errors over the precisely known set x_i (calibration). After building the calibration curve, we make another measurement Y_* at some known x_*.
Then, the confidence interval of the single measurement is given by the expression:

$$\mathbf{Y}_* = \hat{\mathbf{Y}}_* \pm t_{n-p}^{(1/2)\alpha} \cdot S \cdot (1 + u_*)^{1/2} \qquad \text{(I.3), (Seber 5.22)}$$

where

$$S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \cdot (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p} \qquad \text{(Seber § 3.3)}$$

$$u_* = \mathbf{x}_*' (\mathbf{X}' \cdot \mathbf{X})^{-1} \mathbf{x}_* \qquad \text{(Seber 5.17)}$$

$\mathbf{x}_*' = \{1, x_*, ..., x_*^p\}$ – if polynomial curve does not go through

	origin
$\mathbf{x}'_* = \{x_*, \ldots, x_*^p\}$	– if polynomial curve goes through origin
$\hat{\mathbf{Y}}_* = \mathbf{x}'_* \hat{\boldsymbol{\beta}}$	(Seber § 5.2)
n	– number of calibration points
p	– power of the polynomial
t_m^δ	– Student's coefficient at confidence probability $(1-\delta)$ with m degrees of freedom.

Confidence interval for prediction in solving the inverse problem (discrimination)

We build the regression of response values y_i measured with errors at the precisely known set x_i (calibration). After building a calibration curve, we make another measurement Y_*. Using the calibration curve, we find $\hat{x}_*$ - an estimate of the true value x_*:

$$\mathbf{Y}_* = \hat{\mathbf{x}}'_* \hat{\boldsymbol{\beta}}$$

We have to follow the reasoning of (Seber § 7.2.6)

Let $\mathbf{x} = \mathbf{x}_*$ – true value of $\mathbf{x}$ in our measurement.

We define $\hat{\mathbf{Y}}_*$ so that

$$\hat{\mathbf{Y}}_* = \mathbf{x}_* \hat{\boldsymbol{\beta}}$$

Then, from (I.3) we have:

$$\mathbf{Y}_* - \hat{\mathbf{Y}}_* \sim N(0, \sigma^2(1+u_*))$$

and

$$T = \frac{\mathbf{Y}_* - \hat{\mathbf{Y}}_*}{S\sqrt{1+u_*}} \sim t_{n-p}$$

Thus a set of $\mathbf{x}$ which satisfy the condition

$$(\mathbf{Y}_* - \hat{\mathbf{Y}}_x)^2 \le (t_{n-p}^{(1/2)\alpha})^2 \cdot S^2 \cdot (1+u_x) \qquad \text{(I.4)}$$

form a confidence region at $100 \cdot (1-\alpha)$ level for $\mathbf{x}_*$

Here we use:

$$\hat{\mathbf{Y}}_x = \mathbf{x} \cdot \hat{\boldsymbol{\beta}} \qquad \text{(I.5)}$$

$$u_x = \mathbf{x}'(\mathbf{X}' \cdot \mathbf{X})^{-1}\mathbf{x} \qquad \text{(I.6)}$$

In the particular case of linear regression, formula (I.4) is equivalent to (Seber § 7.11).

For linear regression going through origin, formula (I.4) gives result equivalent to (Seber § 7.18).

II. Weighted Regressions

Assumptions, which were used in the previous chapter for constructing the regression

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{(II.1)}$$

are

$\operatorname{cov}[\varepsilon_i, \varepsilon_j] = \delta_{ij} \cdot \sigma^2$

$\mathbf{D}[\varepsilon] = \sigma^2 \cdot \mathbf{I_n}$

Still, following to (Seber § 3.6), we can build a generalized method of least squares.
Let $\mathbf{D}[\varepsilon] = \sigma^2 \cdot \mathbf{V}$, where $\mathbf{V}$ is a known positively-defined matrix of size $(n \times n)$.
In this case, non-singular matrix $\mathbf{K}$ exists so that

$\mathbf{V} = \mathbf{K} \cdot \mathbf{K}'$

We define

$\mathbf{Z} = \mathbf{K}^{-1} \cdot \mathbf{Y}$,

$\mathbf{B} = \mathbf{K}^{-1} \cdot \mathbf{X}$

$\boldsymbol{\eta} = \mathbf{K}^{-1} \cdot \boldsymbol{\varepsilon}$

and build another regression

$$\mathbf{Z} = \mathbf{B} \cdot \boldsymbol{\beta} + \boldsymbol{\eta} \qquad \text{(II.2)}$$

In this regression $E[\boldsymbol{\eta}] = 0$ и $\mathbf{D}[\boldsymbol{\eta}] = \sigma^2 \cdot \mathbf{I_n}$
This means that model (II.2) is equivalent to model (I.1), where all ε_i are not correlated and have the same dispersion.

Least-squires estimate $\hat{\boldsymbol{\beta}}$ for vector $\boldsymbol{\beta}$ is calculated by minimizing of value $\boldsymbol{\eta}' \cdot \boldsymbol{\eta}$ and is given by expression

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}' \cdot \mathbf{B})^{-1} \mathbf{B}' \cdot \mathbf{Z} = (\mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{X})^{-1} \mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{Y} \qquad \text{(II.3)}$$

Just in the same way the expected value $\hat{Y}_*$ at a given known x_* is given by

$\hat{\mathbf{Y}}_* = \mathbf{x}'_* \hat{\boldsymbol{\beta}}$

and true unknown x_* for measured Y_* is estimated by $\hat{x}_*$:

$\mathbf{Y}_* = \hat{\mathbf{x}}'_* \hat{\boldsymbol{\beta}}$

Methods of evaluating confidence intervals are not directly applicable, although (II.2) and (I.1) seems to be equivalent.
We can expect that the statistical behavior of vector $\hat{\boldsymbol{\beta}}$ from (II.3) is analogous to the conventional regression model because (II.3) is the usual estimate made by the least squares method. In the experiment, we are measuring or setting values x_* and Y_*. In general, we cannot match them related values b_* and Z_* in inverted regression.

Fortunately, in some practically significant cases such a match is possible.

We need to make additional assumptions.

Let us assume that errors ε_i are not correlated as previously, that is $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = 0$ when $i \neq j$. Now the dispersions are not the same.
Let us assume that we are making multiple measurements of the detector response at a given strictly known $\tilde{x}$ and by averaging responses we obtain a true response $\tilde{Y}$.
Also, assume that the dispersion of the error depends on $\tilde{x}$ and $\tilde{Y}$ only, that is

$$\tilde{\varepsilon}^2 = \frac{\sigma^2}{w(\tilde{x}, \tilde{Y})} \qquad \text{(II.4)}$$

Also, we have to assume that errors are the same for calibration and for analyte and they follow (II.4).
In practice a particular form of (II.4) is known approximately and is defined by the type of physical experiment.

The most important models are listed below:

$w(x,Y) = 1$ (II.5.1) – conventional regression, no weighting

$w(x,Y) = \frac{1}{|Y|}$ (II.5.2) – error of measurement is proportional to $\sqrt{|Y|}$. For example, this is a case of radioactivity detector.

$w(x,Y) = \frac{1}{Y^2}$ (II.5.3) – Constant relative error? That is $\delta Y \sim |Y|$ and $\frac{\delta Y}{|Y|} = const$

$w(x,Y) = \frac{1}{|x|}$ (II.5.4) – error of measurement is proportional to $\sqrt{|x|}$. Analogous to (II.5.2), if we replace Y with x.

$w(x,Y) = \frac{1}{x^2}$ (II.5.5) – error of measurement is proportional to $|x|$. Analogous to (II.5.3), if we replace Y with x.

We define matrix $\mathbf{V}$ as:

$$\mathbf{V} = diag\{\frac{1}{w(x_i, Y_i)}\} \approx diag\{\frac{1}{w(\tilde{x}_i, \tilde{Y}_i)}\} = \tilde{\mathbf{V}} \qquad \text{(II.6)}$$

where i – is a number of the calibration point.

When we are building the calibration, a precise value $\tilde{Y}_i$ for calibration point is unknown. Therefore, we have to use approximation by replacing $\tilde{Y}_i \to Y_i$. This means that a true error matrix $\tilde{\mathbf{V}}$ is replaced by an approximate $\mathbf{V}$.

An inverse matrix for $\mathbf{V}$ is:

$$\tilde{\mathbf{V}}^{-1} \approx \mathbf{V}^{-1} = diag\{w(\tilde{x}_i, \tilde{Y}_i)\}$$

Confidence interval at $100\cdot(1-\alpha)$ level for the response value at a given x_*.

Using (II.3) we can estimate the true detector response Y_* at a given known x_*. A natural estimate is:

$$\hat{\mathbf{Y}}_* = \mathbf{x}'_*\hat{\boldsymbol{\beta}}$$

We define

$$w_* = w(x_*, \hat{Y}_*) \approx w(x_*, Y_*) = \tilde{w}_*$$

so

$$\varepsilon_*^2 = \frac{\sigma^2}{w(x_*, \hat{Y}_*)} \approx \frac{\sigma^2}{w(x_*, Y_*)} = \tilde{\varepsilon}_*^2$$

is an estimate of dispersion of measured response Y_*, according to (II.4).

Now we define

$$\frac{1}{k_*} = \sqrt{w_*}, \quad Z_* = \frac{Y_*}{k_*}, \quad b_* = \frac{x_*}{k_*},$$

so that

$$\hat{Z}_* = \mathbf{b}'_* \cdot \hat{\boldsymbol{\beta}} = \frac{\mathbf{x}'_*}{k_*} \cdot \hat{\boldsymbol{\beta}} = \frac{\hat{\mathbf{Y}}_*}{k_*}$$

is an estimate for Z_* in inverted regression (II.2).

This means that the method for calculating confidence intervals from chapter for (I) is applicable for Z_* and $\hat{Z}_*$.

Then, a confidence interval for response value of single measurement is given by:

$$Z_* = \hat{Z}_* \pm t_{n-p}^{(1/2)\alpha} \cdot S \cdot (1 + u_*)^{1/2} \qquad \text{(II.7)}$$

Where

$$S^2 = \frac{(\mathbf{Z} - \mathbf{B}\hat{\boldsymbol{\beta}})' \cdot (\mathbf{Z} - \mathbf{B}\hat{\boldsymbol{\beta}})}{n - p}$$

$$u_* = \mathbf{b}'_*(\mathbf{B}' \cdot \mathbf{B})^{-1}\mathbf{b}_*$$

Now, we need to convert from (b, Z) back to (x, Y). We get:

$$Y_* = \hat{Y}_* \pm t_{n-p}^{(1/2)\alpha} \cdot S \cdot \left(\frac{1}{w_*} + U_*\right)^{1/2} \qquad \text{(II.8)}$$

where

$$S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \cdot \mathbf{V}^{-1} \cdot (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p}$$

$$U_* = \mathbf{x}'_*(\mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{X})^{-1}\mathbf{x}_*$$

Confidence interval for prediction in solving an inverse problem (discrimination).

We are making an estimate $\hat{x}_*$ of a true value x_* using the regression (II.3) at a measured detector response Y_*:

$$\mathbf{Y}_* = \hat{\mathbf{x}}_*'\hat{\boldsymbol{\beta}}$$

We need to repeat the reasoning, analogous to the one we made for confidence interval for response value.
In the same way we get:

$$w_* = w(\hat{x}_*, Y_*) \approx w(x_*, Y_*) = \tilde{w}_*$$

$$\frac{1}{k_*} = \sqrt{w_*}\,, \quad Z_* = \frac{Y_*}{k_*}$$

$$\mathbf{Z}_* = \frac{\mathbf{Y}_*}{k_*} = \frac{\hat{\mathbf{x}}_*'}{k_*}\hat{\boldsymbol{\beta}} = \hat{\mathbf{b}}_*\hat{\boldsymbol{\beta}}$$

We define $b_* = \frac{x_*}{k_*}$. An estimate of confidence intervals from chapter (I) is applicable for b_* and $\hat{b}_*$. Namely, a set of all $\mathbf{b}$ which follow the condition

$$(\mathbf{Z}_* - \hat{\mathbf{Z}}_b)^2 \le (t_{n-p}^{(1/2)\alpha} \cdot S)^2 \cdot (1 + u_b) \qquad \text{(II.9)}$$

form a confidence region for $\mathbf{b}_*$ at $100 \cdot (1-\alpha)$ level.
Here we define:

$$\hat{\mathbf{Z}}_b = \mathbf{b} \cdot \hat{\boldsymbol{\beta}}$$

$$u_b = \mathbf{b}'(\mathbf{B}' \cdot \mathbf{B})^{-1}\mathbf{b}$$

Now we need to convert back from (b, Z) to (x, Y). Assuming $\mathbf{b} = \frac{\mathbf{x}}{k_*}$ we get a confidence region for $\mathbf{x}_*$ at $100 \cdot (1-\alpha)$ level. This is a set of all $\mathbf{x}$, which follow the condition:

$$(\mathbf{Y}_* - \hat{\mathbf{Y}}_x)^2 \le (t_{n-p}^{(1/2)\alpha})^2 \cdot S^2 \cdot (\frac{1}{w_*} + U_x) \qquad \text{(II.10)}$$

where

$$S^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \cdot \mathbf{V}^{-1} \cdot (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-p}$$

$$U_x = \mathbf{x}' \cdot (\mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{X})^{-1} \cdot \mathbf{x}$$

$$\hat{\mathbf{Y}}_{\mathbf{x}} = \mathbf{x} \cdot \hat{\boldsymbol{\beta}}$$

III. Examples

The examples below represent typical regressions of different polynomial powers and different weighting models. For demonstration purposes, the simulated data are generated with significant error. Confidence intervals for response values at 0.95 confidence probability are drawn.

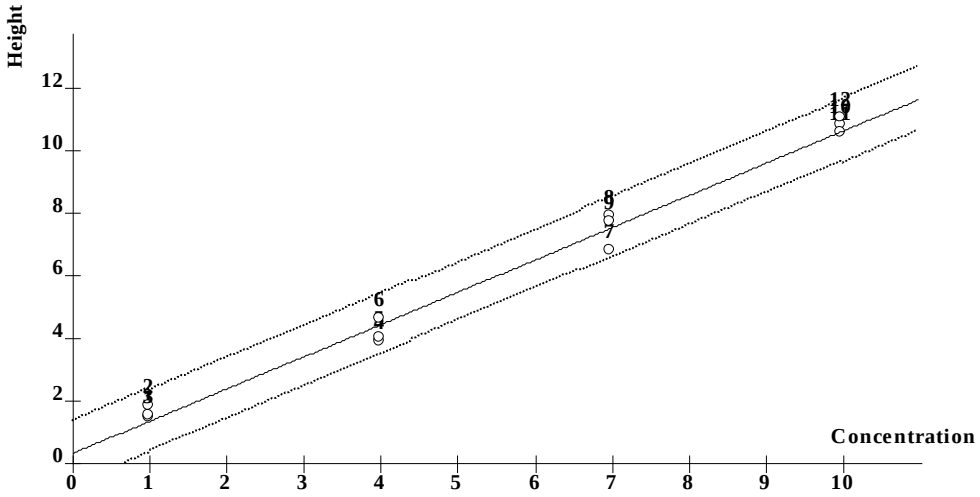

Figure 1. Linear polynomial not going through the origin. The dispersion of error is the same for all measurements, no weighting.

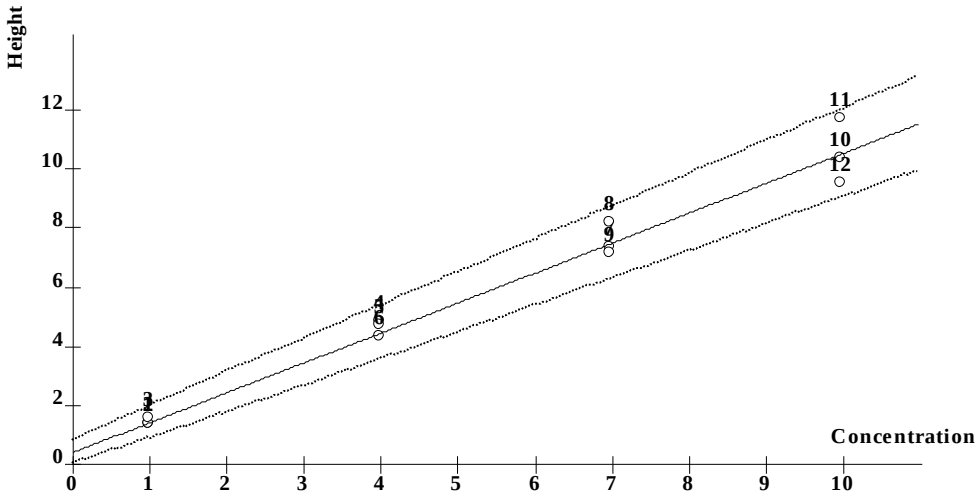

Figure 2. Linear polynomial not going through the origin. The dispersion of error is proportional to $\sqrt{Height}$. Weighting 1/*Height*.

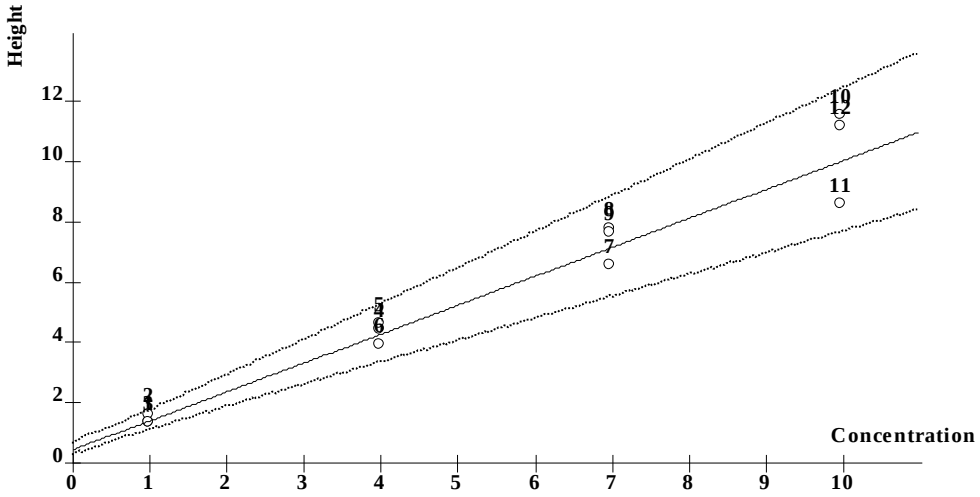

Figure 3. Linear polynomial not going through the origin. The dispersion of error is proportional to *Height* (constant relative error). Weighting 1/*Height*^{2}.

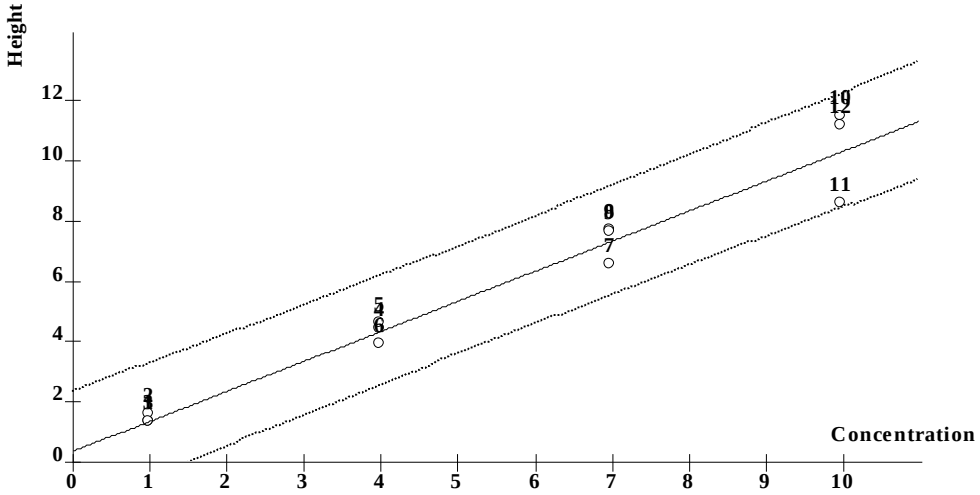

Figure 4. Linear polynomial not going through the origin, the same data as on Fig.3. The dispersion of error is proportional to *Height* (constant relative error). Regression is constructed without weighting (just as if absolute error is constant). This regression gives an approximation formula that is not quite correct and has incorrect confidence intervals.

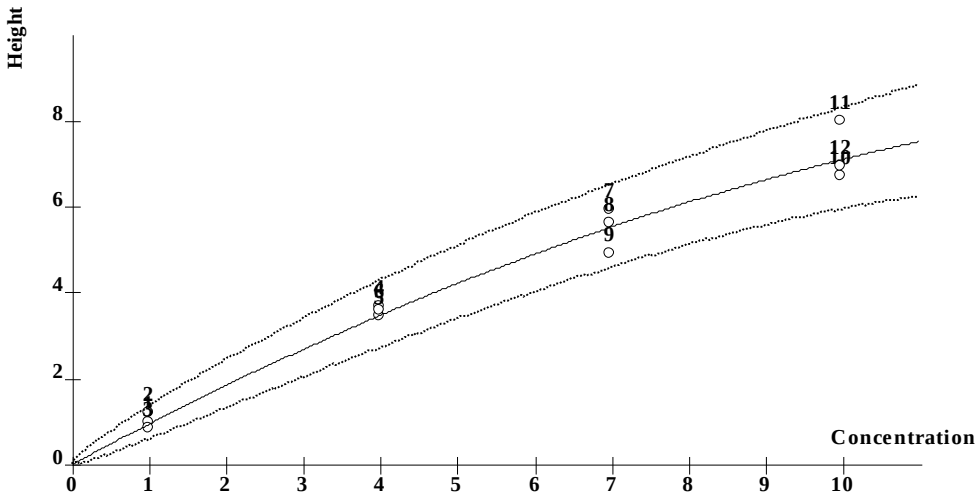

Figure 5. Quadratic polynomial through the origin. The dispersion of error is proportional to $\sqrt{Height}$. Weighting 1/*Height*.

A correct test plan for calibration is especially important when building regression with a non-linear polynomial. Let us consider a regression of the simulated data which should be approximated by quadratic polynomial. If we select 3 concentrations and make measurements twice at each concentration, the calibration curve could look quite nice. Still, when we calculate and draw related confidence intervals we would notice that adequate results are possible near the initial concentrations only. Therefore, this test plan is incorrect.

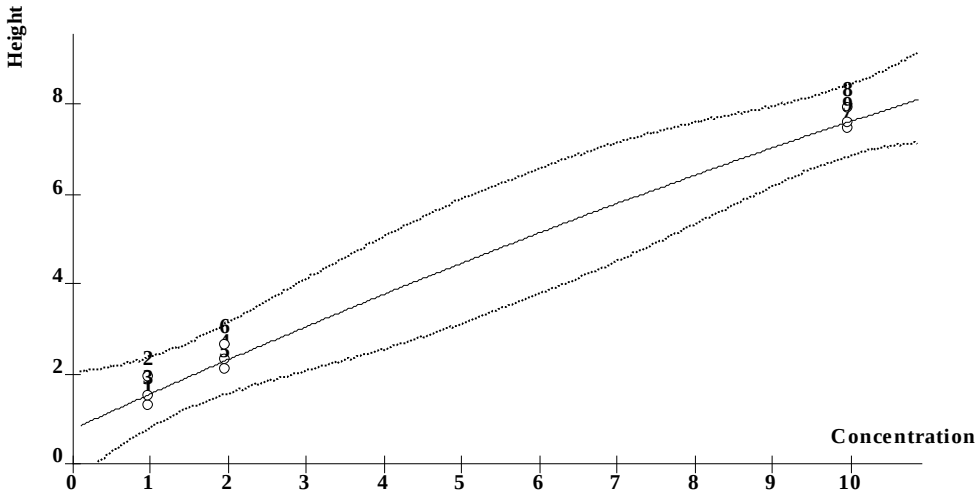

Figure 6. Quadratic polynomial not going through origin. The dispersion of error is the same for all measurements, no weighting. The simulated detector response is measured three times at three concentrations. This test plan produces a poor prediction.

A prediction becomes much better if we select uniformly distributed concentrations for the calibration. This calibration curve would supply good predictions over an entire range of calibration.

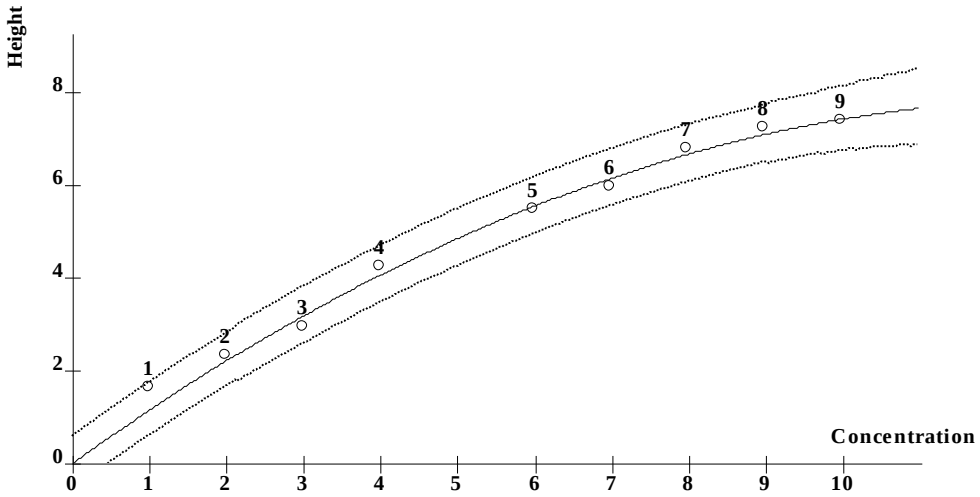

Figure 7. Quadratic polynomial not going through origin. The errors are the same for all measurements, no weighting. The concentrations for calibration are uniformly distributed. This test plan produces an adequate prediction.

IV.Conclusion

We have strict mathematical expressions for confidence intervals for detector response and for the prediction in an inverse problem in the case of generalized regression with weighting.
An expressions (II.8) and (II.10) are generalizations of (I.3) and (I.4) respectively. They are applicable either for linear regressions or for regressions with polynomial of other powers (quadratic, cubic etc.).

References

1. G. Seber Linear Regression Analysis