

# Regression methods

Regression methods search for the best relationship between a set of variables describing objects (samples) and a set of responses obtained for the same objects.

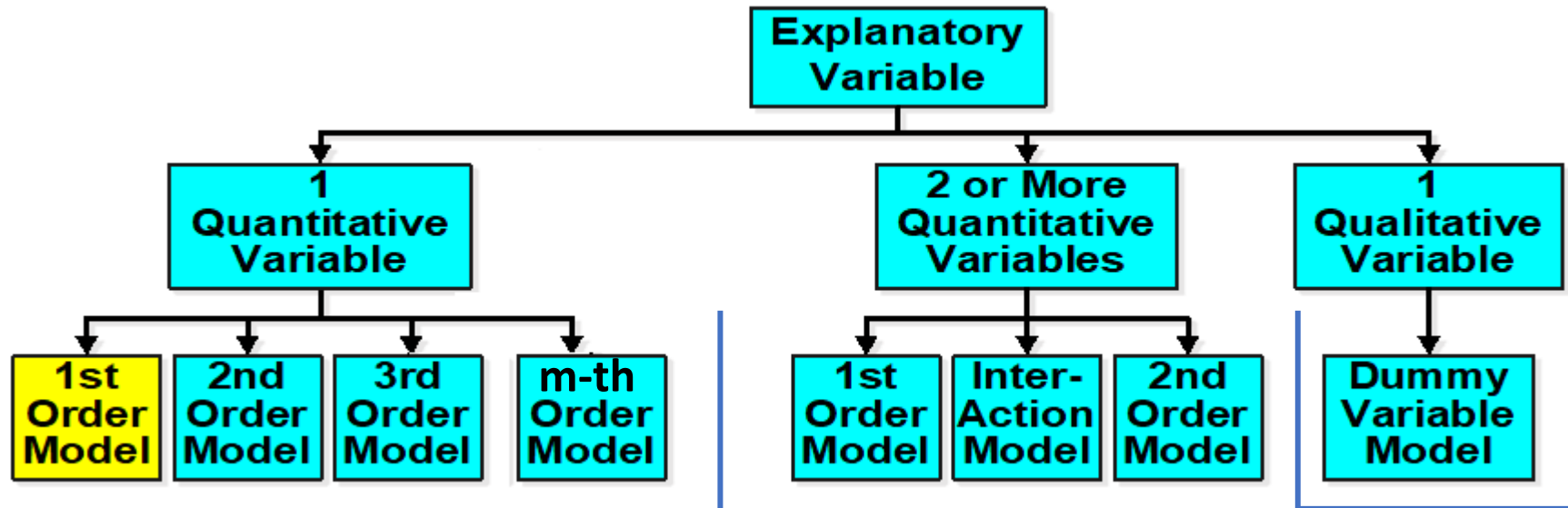
The type of relationship (model) is initially searched for by fitting it to the experimental measures; once the model has been subjected to validation, i.e., the quality of its description of responses is checked, a prediction of future responses can be made:



Regression methods are thus mathematical models searching for functions  $f$  that relate to a response  $Y$  a certain number,  $m$ , of variables (descriptors):

$$Y = f(X_1, X_2, \dots, X_m)$$

As shown in the general scheme reported in the next slide, the number and type of variables determine the specific definition of a regression model.



### Univariate models

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

.....

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_m X^m + \varepsilon$$

### Bivariate models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \varepsilon$$

### Multivariate models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon$$

## Linear regression

A linear model of type:

$$Y = \beta_0 + \sum_{i=1}^m \beta_i X_i$$

with  $\beta_0, \beta_1, \dots, \beta_m$  representing the model parameters, is the simplest type of regression model.

Actually, even a generic transformation of each variable,  $f_i(X_i)$  can be used in the model:

$$Y = \beta_0 + \sum_{i=1}^m \beta_i f_i(X_i)$$

Examples of linear models are then:

$$Y = \beta_0 + \beta_1 \log X \qquad Y = \beta_0 + \beta_1 \frac{1}{X_1} + \beta_2 \sinh(X_2)$$

whereas the following are examples of non linear models:

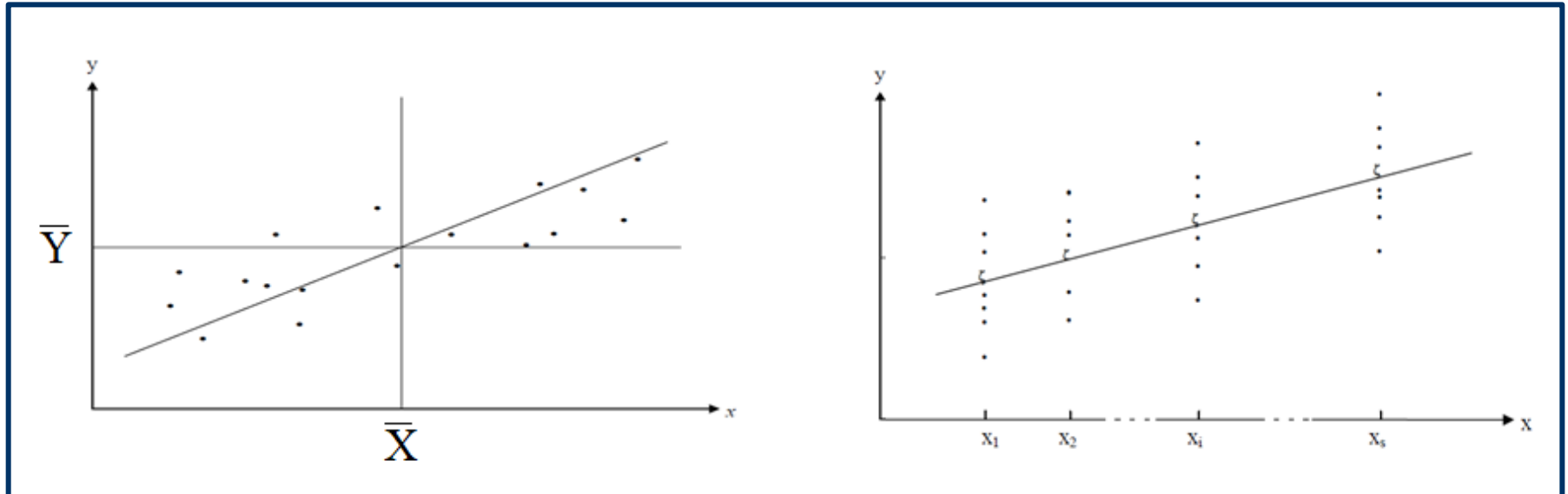
$$Y = \beta_1 \exp(\beta_2 X) \qquad Y = \beta_1 \exp(\beta_2 (X - \beta_3)^2)$$

## Simple linear regression

A linear model of type:  $Y = \beta_0 + \beta_1 X + \varepsilon$

thus involving a single variable, represents simple linear regression, that is one of the most common in analytical chemistry (for example, it is used for calibration purposes).

Examples of sets of observations that can be treated with simple linear regression are:



where average values of X and Y, representing the center of gravity (or centroid) of the measurements set, are also indicated in the left graph, whereas the case of replicated measurements of response y for specific values of the independent variable (x) is represented in the right graph.

A **deterministic** and an **accidental (or stochastic) component** can be distinguished in a simple linear model:

$$Y = \underbrace{\beta_o + \beta_1 X}_{\text{Deterministic component}} + \varepsilon$$

*Accidental (stochastic) component*  
↓

For the **i-th observation** the equation becomes:

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

The **main assumptions** made for simple linear regression are:

$\varepsilon_i \sim N(0, \sigma^2)$ , thus:  $E(\varepsilon_i) = 0$  and  $V(\varepsilon_i) = \sigma^2$  for  $i = 1, 2, \dots, n$  (*homoscedasticity*);

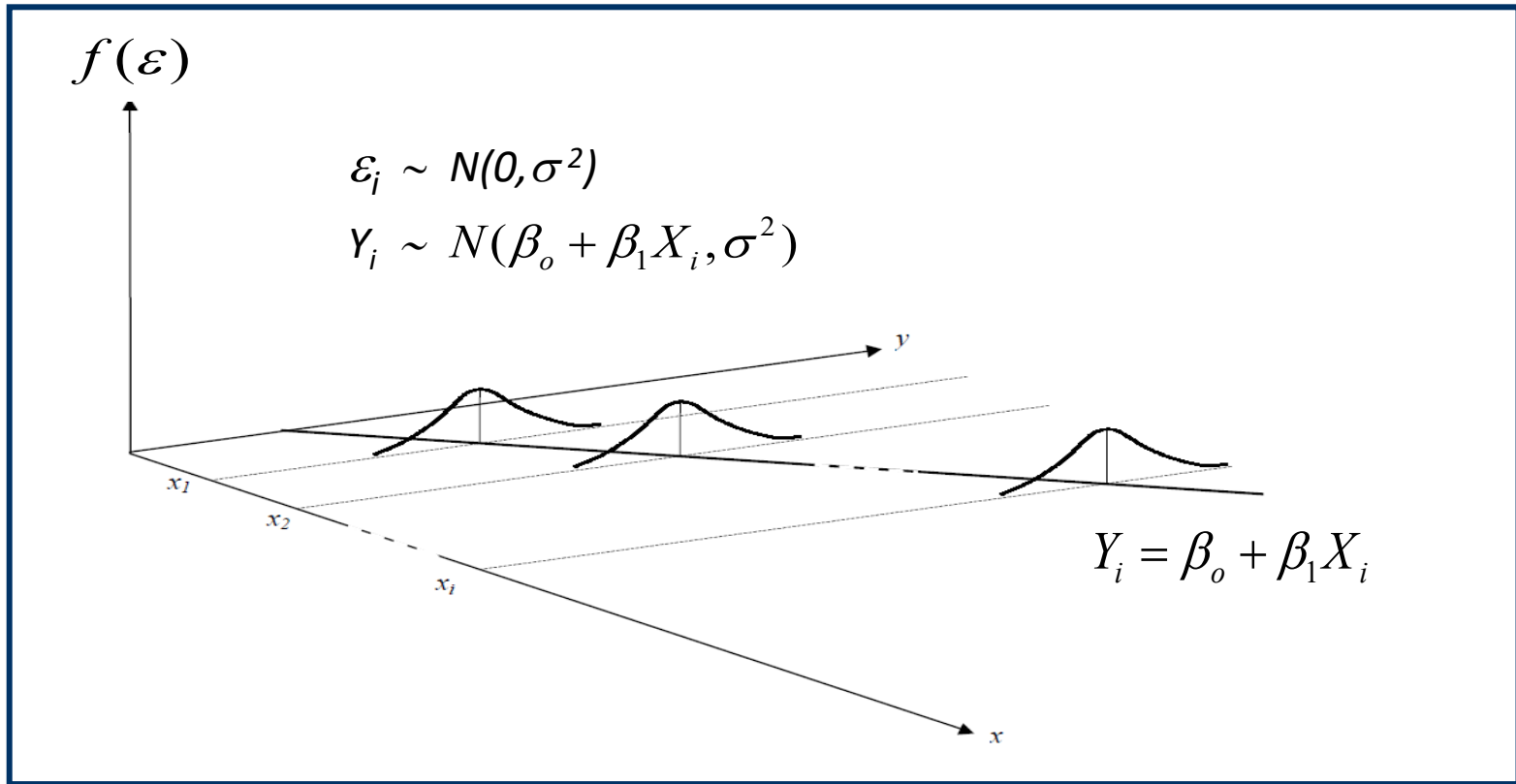
$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for  $i \neq j = 1, 2, \dots, n$  (*uncorrelation between stochastic components*)

Consequently:

$$E(Y_i) = \beta_o + \beta_1 X_i \qquad Y_i \sim N(\beta_o + \beta_1 X_i, \sigma^2) \quad \text{for } i = 1, 2, \dots, n;$$

$$V(Y_i) = \sigma^2 \quad \text{for } i = 1, 2, \dots, n; \qquad \text{Cov}(Y_i, Y_j) = 0 \quad \text{for } i \neq j = 1, 2, \dots, n.$$

Such assumptions can be represented with the following graph:



Parameters  $\beta_0$  and  $\beta_1$  of the model can be subjected to **statistical inference**.

One of the first approaches adopted at this aim, still one of the most common, is based on the **least squares method**, that was described in 1805 by the French mathematician Adrien-Marie Legendre.

**Friedrich Gauss** was also one of the first mathematicians to study linear regression.

Estimators of  $\beta_0$  and  $\beta_1$  are indicated as  $b_0$  and  $b_1$  and the general equation for values predicted by the model is:

$$\hat{Y} = b_0 + b_1 X$$

Minimization of the squares of differences between actual and modelled values leads to the following equations for the two estimators:

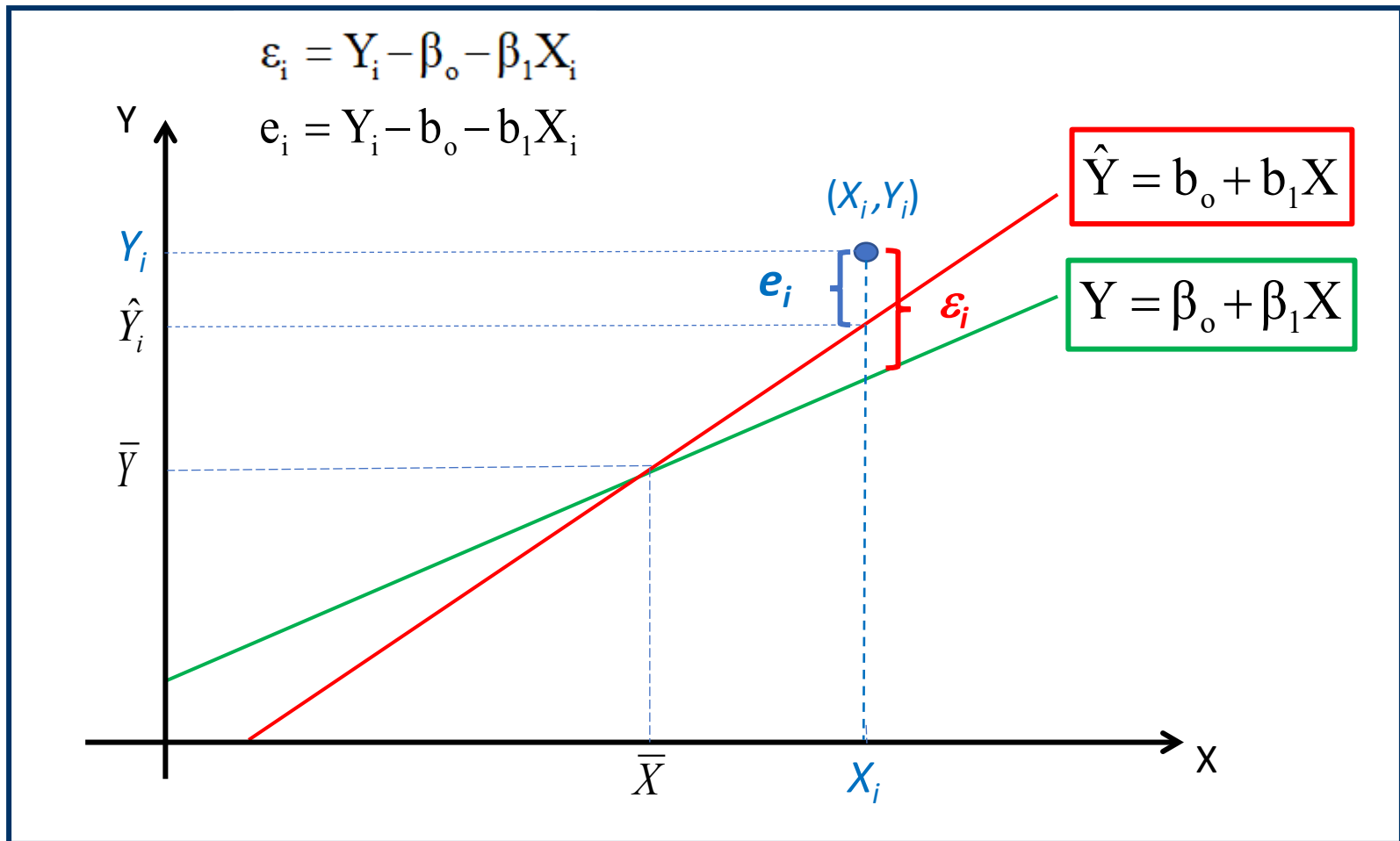
$$\left\{ \begin{array}{l} b_0 = \bar{Y} - b_1 \bar{X} \\ b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} \end{array} \right.$$

Combining the equation for  $b_0$  and the general equation of the model, the following equation is easily obtained:

$$\hat{Y} = \bar{Y} + b_1 (X - \bar{X})$$

It is thus apparent that the least squares regression line passes through the data centroid, as emphasized in the following graph.

Given a specific couples of values  $(X_i, Y_i)$  the meaning of  $\varepsilon_i$  as the difference between the experimental value  $Y_i$  and the value provided by the real model (green line) is evidenced in the figure:

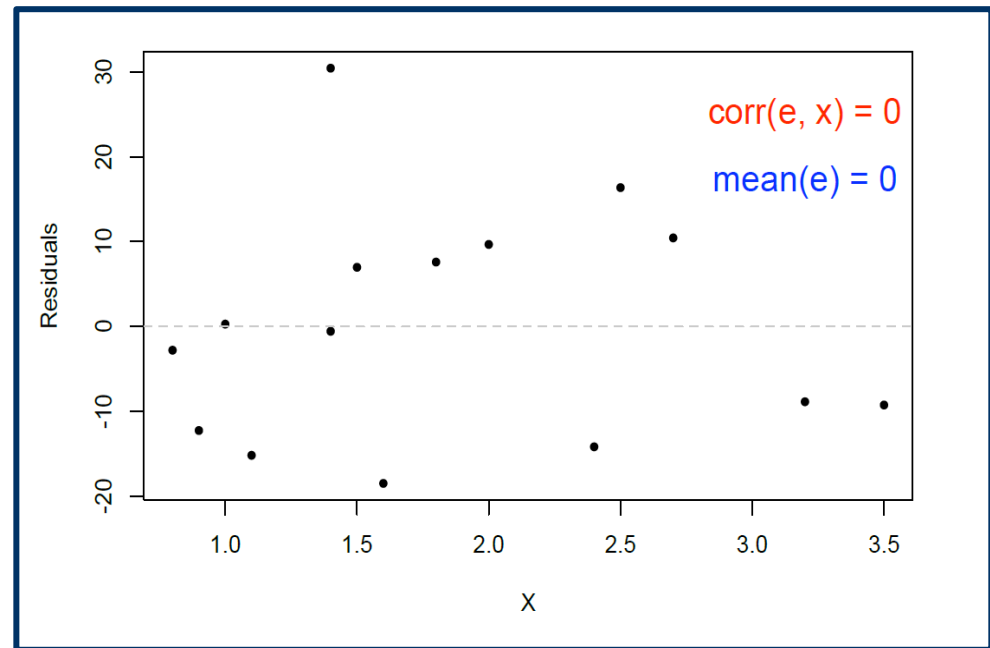


On the other hand, the difference between the experimental value  $Y_i$  and the response value predicted by the inferred model (red line) is represented by  $e_i$ , which is called residual.



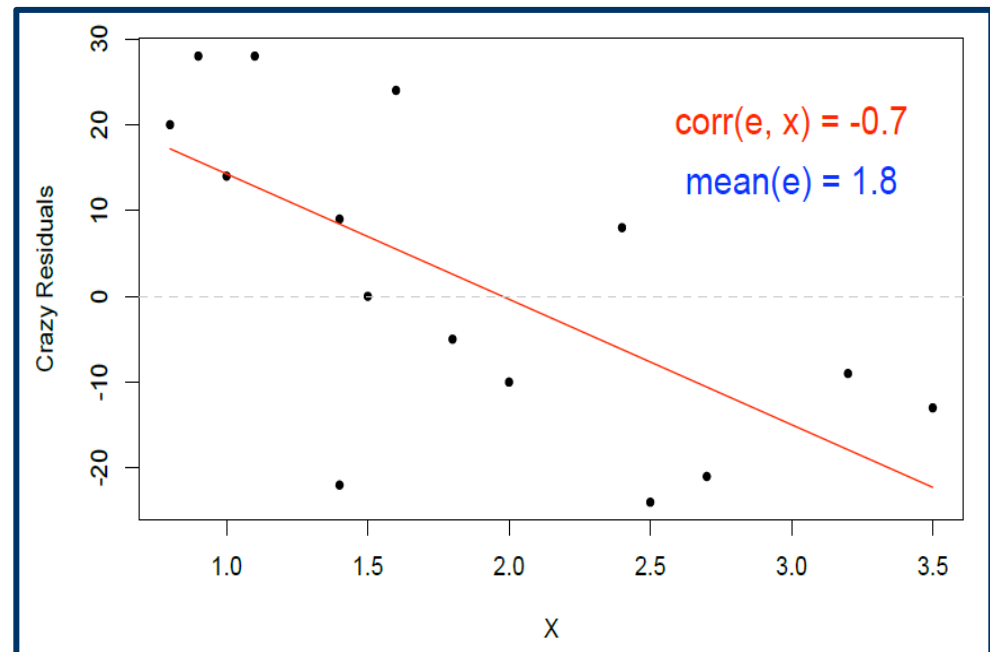
If experimental data are fitted properly, the mean of residuals  $e_i$  and their correlation with  $x$  are both equal to 0:

Actually, also the sum of all residuals is equal to 0.



On the contrary, in the example shown on the right the model is wrong; thus, both the residuals mean and their correlation with  $x$  are not equal to 0:

Specifically, the correlation is negative because the residuals decrease (even becoming negative) at the increase of  $x$ .



Interestingly, equations for estimators  $b_0$  and  $b_1$  can be obtained starting from the assumptions that the mean of residuals (or their sum) and their correlation with  $X$  are equal to 0:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n e_i = 0 &\Rightarrow \frac{1}{n} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \\ &\Rightarrow \bar{Y} - b_0 - b_1 \bar{X} = 0 \\ &\Rightarrow b_0 = \bar{Y} - b_1 \bar{X}\end{aligned}$$

$$\begin{aligned}\text{corr}(e, X) = 0 &\Rightarrow \sum_{i=1}^n e_i (X_i - \bar{X}) = 0 \\ &\Rightarrow \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)(X_i - \bar{X}) = 0 \\ &\Rightarrow \sum_{i=1}^n (Y_i - \bar{Y} - b_1 (X_i - \bar{X}))(X_i - \bar{X}) = 0 \\ &\Rightarrow b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \Leftrightarrow b_1 = \frac{S_{XY}}{S_{XX}}\end{aligned}$$

## Variances and distributions for estimators $b_0$ and $b_1$

Variances for the **estimators of regression line slope and intercept** can be obtained considering the **general properties of variance**:

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i Y_i - \bar{X} Y_i - \cancel{X_i \bar{Y}} + \cancel{\bar{X} Y_i})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n a_i Y_i$$
$$\text{where: } a_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Consequently:

$$V(b_1) = V\left[\sum_{i=1}^n a_i Y_i\right] = \sum_{i=1}^n a_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \left[ \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$= \sigma^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sigma^2}{S_{XX}}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum_{i=1}^n Y_i}{n} - \frac{\left[ \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \bar{X}} = \sum_{i=1}^n \left[ \frac{1}{n} - \frac{(X_i - \bar{X}) \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] Y_i = \sum_{i=1}^n a_i Y_i$$

Note that  $a_i$  in this case is different from the one used for  $V(b_1)$ .



$$V(b_0) = \sum_{i=1}^n a_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \sum_{i=1}^n \left[ \frac{1}{n} - \frac{(X_i - \bar{X}) \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]^2$$

When each square of the binomial indicated in the last member of the sequence of equations is calculated and then the sum of squares is considered, the following equations have to be taken into account for cross-products:

$$\sum_{i=1}^n \left[ -\frac{2(X_i - \bar{X}) \bar{X}}{n \sum_{i=1}^n (X_i - \bar{X})^2} \right] = -\frac{2\bar{X}}{n \sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X}) = -\frac{2\bar{X}}{n \sum_{i=1}^n (X_i - \bar{X})^2} \left[ \sum_{i=1}^n X_i - n \bar{X} \right] = 0$$

Consequently:

$$V(b_0) = \sum_{i=1}^n a_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \left[ \sum_{i=1}^n \frac{1}{n^2} + \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \bar{X}^2}{\left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2} \right]$$

Since:  $\sum_{i=1}^n \frac{1}{n^2} = n \frac{1}{n^2} = \frac{1}{n}$

the **variance of  $b_0$**  can be finally expressed as:

$$V(b_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)$$

The properties of  $b_0$  and  $b_1$  can thus be summarized as follows:

$$\begin{aligned} E(b_0) &= \beta_0 & E(b_1) &= \beta_1 \\ V(b_0) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right) & V(b_1) &= \frac{\sigma^2}{S_{XX}} \end{aligned}$$

It is worth noting that  $\sigma^2$  can be estimated using the residual mean square,  $MS_{\text{RES}}$ , corresponding to the square of residual standard deviation,  $s_{y/x}$ .

Distributions for  $b_0$  and  $b_1$  can thus be obtained:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$b_0 \sim N\left(\beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)\right)$$

$$\frac{b_0 - \beta_0}{\sqrt{s_{y/x}^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \sim t_{n-2}$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

$$\frac{b_1 - \beta_1}{\sqrt{s_{y/x}^2 / S_{XX}}} \sim t_{n-2}$$

Confidence intervals for  $b_0$  and  $b_1$  can then be expressed as follows:

$$b_0 \pm t_{n-2, 1-\alpha/2} \sqrt{s_{y/x}^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)} \qquad b_1 \pm t_{n-2, 1-\alpha/2} \sqrt{\frac{s_{y/x}^2}{S_{XX}}}$$

A confidence interval can be also calculated for the predicted response  $Y$ , according to the linear regression model, once a specific value of variable  $X$  is fixed.

First, the expectation for the predicted response can be expressed as:

$$E(\hat{Y}|X) = E(b_0 + b_1 X) = E(b_0) + X E(b_1) = \beta_0 + \beta_1 X$$

Since estimators  $b_0$  and  $b_1$  are NOT independent, the following equation has to be written for the predicted response variance:

$$\begin{aligned} V(\hat{Y}) &= V(b_0 + b_1 X) = V(b_0) + V(b_1 X) + 2 \text{COV}(b_0, b_1 X) \\ &= V(b_0) + X^2 V(b_1) + 2 X \text{COV}(b_0, b_1) \end{aligned}$$

since  $X$ , in this case, is a specific value (thus it is a constant).

Covariance between  $b_0$  and  $b_1$  can be calculated starting from the following general property:

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

which implies that:

$$\text{Cov}(aY, bY) = ab \text{Cov}(Y, Y) = ab V(Y)$$

As shown before, both  $b_0$  and  $b_1$  can be expressed as  $\sum_{i=1}^n a_i Y_i$

with  $a_i$  having a different expression in the two cases:

$$\left[ \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right] \text{ for } b_0 \quad \left[ \frac{(X_i - \bar{X})}{S_{XX}} \right] \text{ for } b_1$$

Consequently:

$$\begin{aligned} \text{Cov}(b_0, b_1) &= \text{Cov} \left[ \sum_{i=1}^n \left( \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) Y_i, \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{S_{XX}} \right) Y_i \right] = \sum_{i=1}^n \text{Cov} \left[ \left( \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) Y_i, \left( \frac{X_i - \bar{X}}{S_{XX}} \right) Y_i \right] = \\ &= \sum_{i=1}^n \left( \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \left( \frac{X_i - \bar{X}}{S_{XX}} \right) \text{Cov}(Y_i, Y_i) = \sum_{i=1}^n \left( \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \left( \frac{X_i - \bar{X}}{S_{XX}} \right) V(Y_i) \end{aligned}$$



Considering that  $V(Y_i)$  is always equal to  $\sigma^2$ , if homoschedasticity occurs, the following equation can be obtained:

$$\begin{aligned} \text{Cov}(b_0, b_1) &= \sigma^2 \sum_{i=1}^n \left[ \left( \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \left( \frac{X_i - \bar{X}}{S_{XX}} \right) \right] = \sigma^2 \sum_{i=1}^n \left[ \frac{(X_i - \bar{X})}{nS_{XX}} - \frac{(X_i - \bar{X})^2 \bar{X}}{S_{XX}^2} \right] = \\ &= \sigma^2 \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})}{nS_{XX}} - \frac{\bar{X}S_{XX}}{S_{XX}^2} \right] = \sigma^2 \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})}{nS_{XX}} - \frac{\bar{X}}{S_{XX}} \right] = \sigma^2 \left[ \frac{n\bar{X} - n\bar{X}}{nS_{XX}} - \frac{n\bar{X}}{nS_{XX}} \right] = -\sigma^2 \frac{\bar{X}}{S_{XX}} \end{aligned}$$

Considering the expressions for  $V(b_0)$ ,  $V(b_1)$  and  $\text{Cov}(b_0, b_1)$ :

$$V(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right) \quad V(b_1) = \frac{\sigma^2}{S_{XX}} \quad \text{Cov}(b_0, b_1) = -\sigma^2 \frac{\bar{X}}{S_{XX}}$$

the variance related to the predicted response can be finally obtained:

$$\begin{aligned} V(\hat{Y}) &= V(b_0) + X^2 V(b_1) + 2X \text{COV}(b_0, b_1) = \\ &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right) + X^2 \frac{\sigma^2}{S_{XX}} + 2X \left( -\frac{\bar{X}\sigma^2}{S_{XX}} \right) = \frac{\sigma^2}{n} + \sigma^2 \frac{\bar{X}^2}{S_{XX}} + \sigma^2 \frac{X^2}{S_{XX}} - \sigma^2 \frac{2X\bar{X}}{S_{XX}} \end{aligned}$$

The equation can be easily re-written in a compact form:

$$V(\hat{Y}) = \sigma^2 \left( \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}} \right)$$

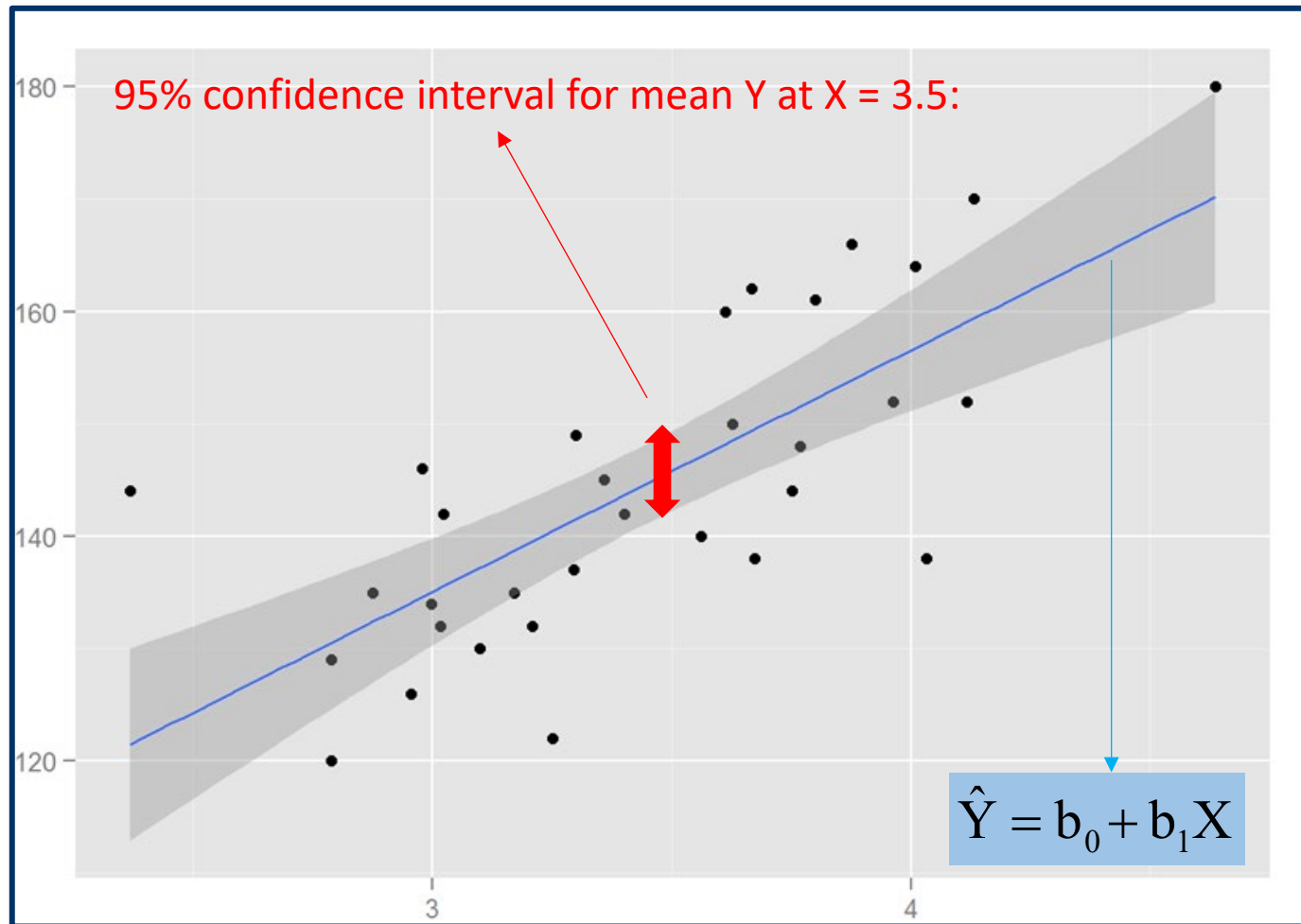
The following expression can subsequently be written, considering the normality of distribution for predicted Y values:

$$\frac{\hat{Y} - E(\hat{Y})}{[V(\hat{Y})]^{1/2}} \sim t_{n-2}$$

Using  $s_{y/x}$  as an estimator for  $\sigma$ , the confidence interval for the predicted values of Y can be expressed as:

$$(b_0 + b_1 X) \pm t_{n-2, 1-\alpha/2} S_{Y/X} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}$$

This expression is exploited by fitting programs to draw the so-called confidence bands, together with regression line, as shown, for example, in the next figure.

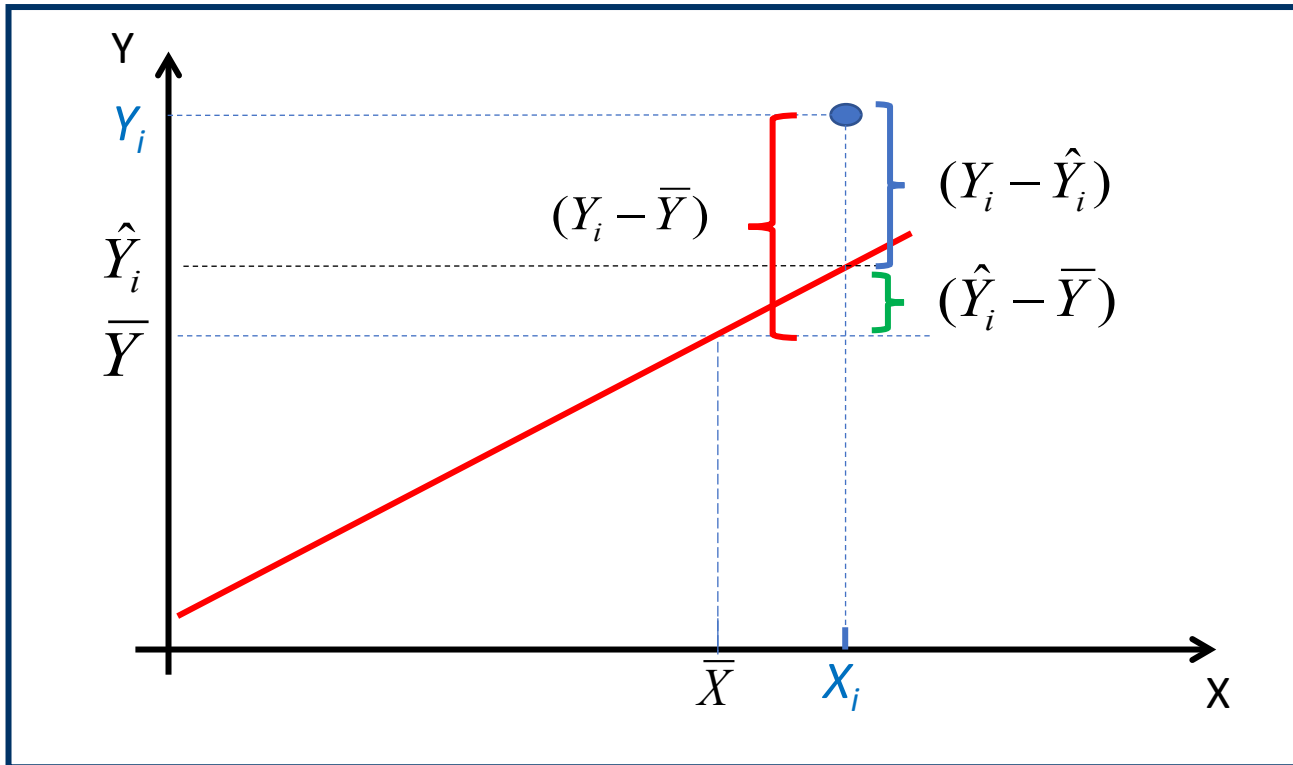


As apparent, the narrowest confidence interval is observed when an X value close to the system centroid is considered, as clearly indicated by the last equation.

On the other hand, it seems that several actual responses can be located outside confidence bands. This result will be discussed later.

## Partitioning of total deviations and ANOVA related to linear regression

As shown in the following figure, the difference between an experimental Y value and the average of Y values can be partitioned into two contributions:



$$\boxed{Y_i - \bar{Y}} = Y_i - \bar{Y} + \hat{Y}_i - \hat{Y}_i = \boxed{(\hat{Y}_i - \bar{Y})} + \boxed{(Y_i - \hat{Y}_i)}$$

A further development of this equation is:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i [(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

Indeed, the double product  $2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$  is equal to zero:

$$\begin{aligned} Y_i - \hat{Y}_i &= Y_i - (b_0 + b_1 X_i) \\ &= Y_i - (\bar{Y} - b_1 \bar{X} + b_1 X_i) \\ &= (Y_i - \bar{Y}) - b_1 (X_i - \bar{X}) \end{aligned}$$

$$\begin{aligned} \hat{Y}_i - \bar{Y} &= (b_0 + b_1 X_i) - \bar{Y} \\ &= (\bar{Y} - b_1 \bar{X}) + b_1 X_i - \bar{Y} \\ &= b_1 (X_i - \bar{X}) \end{aligned}$$



$$\begin{aligned} \sum_{i=1}^n [(Y_i - \bar{Y}) - b_1 (X_i - \bar{X})] \times b_1 (X_i - \bar{X}) &= \sum_{i=1}^n [b_1 (X_i - \bar{X})(Y_i - \bar{Y}) - b_1^2 (X_i - \bar{X})^2] = \\ &= b_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) - b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 = b_1 S_{XY} - b_1^2 S_{XX} = b_1 (S_{XY} - b_1 S_{XX}) = \\ &= b_1 \left( S_{XY} - \frac{S_{XY}}{S_{XX}} S_{XX} \right) = b_1 (S_{XY} - S_{XY}) = 0 \end{aligned}$$

The total deviation of a specific value from the mean of values can thus be partitioned into the regression deviation and the residual deviation:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

**Total deviation**    **Regression deviation**    **Residual deviation**



$$SS_{TOT} = SS_{REG} + SS_{RES}$$

The strict analogy with the equation referred to one-way (fixed factor) ANOVA is clear.

The following ANOVA table for regression can thus be written:

Source of variation	SS	df	MS	E(MS)
Regression	$\sum_i (\hat{Y}_i - \bar{Y})^2$	1	$MS_{REG}$	$\sigma^2 + \beta_1^2 S_{XX}$
Residuals	$\sum_i (Y_i - \hat{Y}_i)^2$	$n-2$	$MS_{RES}$	$\sigma^2$
Total	$\sum_i (Y_i - \bar{Y})^2$	$n-1$		

The expected value for the regression mean square,  $MS_{REG}$ , is identical to the one for the regression sum of squares  $SS_{REG}$  (since  $df = 1$ ).  $SS_{REG}$  can be easily calculated:

$$\begin{aligned}
 SS_{REG} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 X_i - \bar{Y})^2 = \sum_i (\bar{Y} - b_1 \bar{X} + b_1 X_i - \bar{Y})^2 = \\
 &= \sum_i (b_1 X_i - b_1 \bar{X})^2 = b_1^2 \sum_i (X_i - \bar{X})^2 = b_1^2 S_{XX}
 \end{aligned}$$



$$E(SS_{REG}) = S_{XX} E(b_1^2)$$

The **expected value for the square of  $b_1$**  can be calculated starting from its variance and from an already described general property of variance:

$$\text{Var}(b_1) = E(b_1^2) - (E(b_1))^2 \quad \Rightarrow \quad E(b_1^2) = \text{Var}(b_1) + (E(b_1))^2$$

$$\text{Var}(b_1) = \frac{\sigma^2}{S_{XX}} \quad (E(b_1))^2 = \beta_1^2 \quad \Rightarrow \quad E(b_1^2) = \frac{\sigma^2}{S_{XX}} + \beta_1^2$$

Considering that  **$df_{\text{REG}} = 1$ , thus  $MS_{\text{REG}} = SS_{\text{REG}}$** , the following equations can be written:

$$E(MS_{\text{REG}}) = E(SS_{\text{REG}}) = S_{XX} E(b_1^2) = S_{XX} \left[ \frac{\sigma^2}{S_{XX}} + \beta_1^2 \right] = \sigma^2 + \beta_1^2 S_{XX}$$



The significance of regression can be tested using hypothesis testing:

$$\left[ \begin{array}{l} H_0: \beta_1 = 0 \text{ *regression not significant*} \\ H_1: \beta_1 \neq 0 \text{ *regression significant*} \end{array} \right] \Rightarrow \left[ \begin{array}{l} Y = \beta_0 \\ Y = \beta_0 + \beta_1 X \end{array} \right]$$

Since  $E(MS_{REG}) = \sigma^2 + \beta_1^2 S_{XX}$

if the  $H_0$  hypothesis is true, both  $MS_{REG}$  and  $MS_{RES}$  are estimators of  $\sigma^2$ .

The reject criterion for  $H_0$  is thus:

$$\frac{MS_{REG}}{MS_{RES}} \geq F_{\alpha(1, n-2)}$$

## Coefficient of determination and correlation coefficient

The coefficient of determination, denoted as  $R^2$  or  $r^2$ , is the proportion of the variance in the dependent variable that is predictable from the independent variable.

In the case of linear regression, the mathematical definition is:

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = 1 - \frac{SS_{\text{RES}}}{SS_{\text{TOT}}}$$

$R^2$  is thus comprised between 0 and 1.

Notably,  $R^2 = 0$  when  $SS_{\text{REG}} = 0$  and  $R^2 = 1$  when  $SS_{\text{RES}} = 0$ , i.e., when all the observed points are perfectly located on the regression line (total fit of the adopted model).

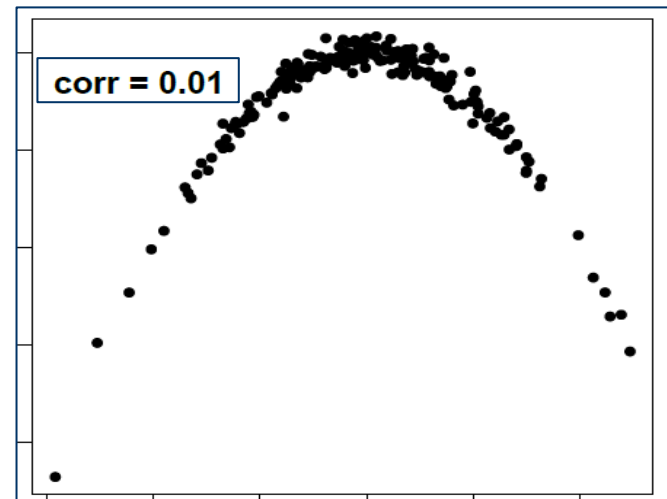
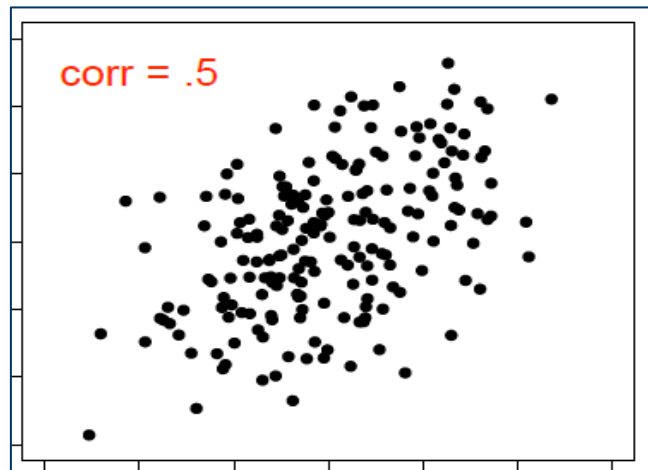
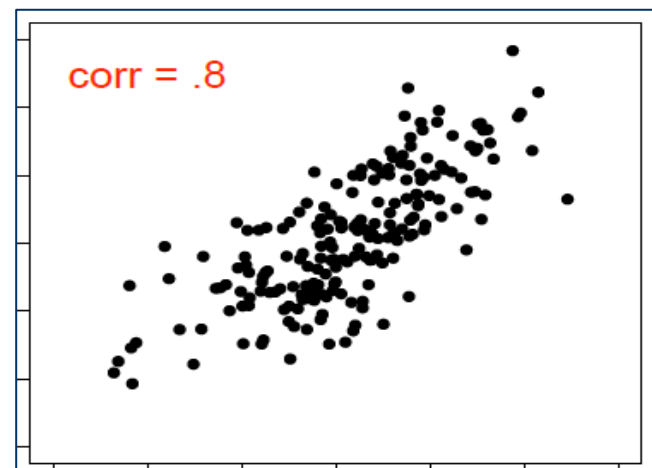
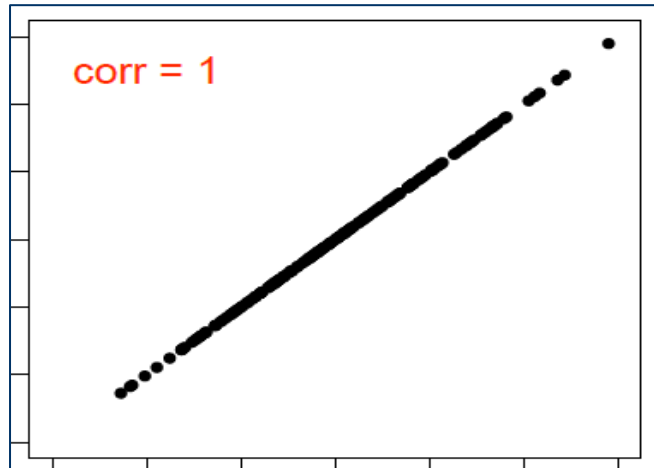
An interesting relationship can be found between the coefficient of determination and the linear correlation coefficient,  $\rho^2_{XY}$  (whose estimator is indicated as  $r^2_{XY}$ ):

$$SS_{\text{REG}} = b_1^2 \sum_i (X_i - \bar{X})^2 = \frac{[\sum_i (X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\sum_i (X_i - \bar{X})^2]^2} \sum_i (X_i - \bar{X})^2 = \frac{(n-1)^2 [\text{Cov}(X,Y)]^2}{(n-1) V(X)} = (n-1) \frac{[\text{Cov}(X,Y)]^2}{V(X)}$$

thus:

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = (n-1) \frac{[\text{Cov}(X,Y)]^2}{V(X)} \frac{1}{(n-1)V(Y)} = \frac{[\text{Cov}(X,Y)]^2}{V(X) V(Y)} = \rho^2_{XY}$$

As shown in the following figure,  $R^2$  becomes much lower than 1 when the variability of data not explained by regression becomes remarkable, yet its approach to values close to 0 does not necessarily mean that variables are not related at all:



Since:  $R^2 = \frac{SS_{REG}}{SS_{TOT}} = \frac{SS_{REG}}{S_{YY}} = \frac{b_1^2 S_{XX}}{S_{YY}}$

and:  $S_{YY} = (n-1)\sigma^2$

$R^2$  is expected to increase when  $S_{xx}$  is increased (i.e., x values are more spread out around their mean) and if  $\sigma^2$  is decreased.

## Comparison between confidence and prediction intervals

When least squares regression is applied, a general model is assumed to be valid for Y:

$$Y = \beta_0 + \beta_1 X \quad \Rightarrow \quad E(Y|X = x) = \beta_0 + \beta_1 x$$

Regression produces an **estimate of model parameters**, thus leading to the equation:

$$Y = b_0 + b_1 X$$

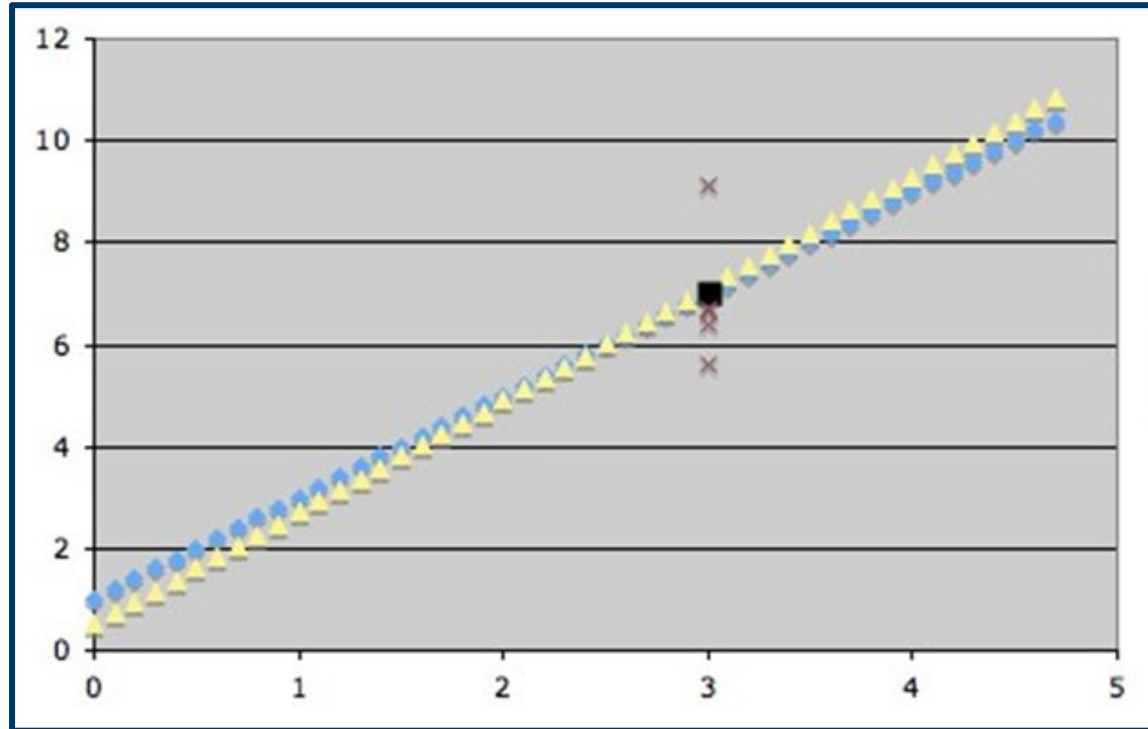
Given **a particular value of X, x**, an estimate of the expected value of Y (the latter is also called **conditional mean**) is obtained:

$$\hat{Y} = b_0 + b_1 x$$

It is important to point out that **the latter is an estimate of the mean, not of a single actual value obtained for Y when X = x.**

**A significant difference may exist between the estimate of conditional mean and that of an actual value of Y.**

In the following figure, blue points correspond to the actual line of conditional means, i.e., the values of the true model, whereas yellow points represent the calculated regression line, that has been found, as it always occur with regression procedures, by rotation around the centroid:



The black square shows the value of the (conditional) mean of Y obtained when  $x = 3$ , which is very close to the true mean.

However, Y values obtained for  $x = 3$ , represented by brown crosses, can be quite different from that value.

In order to obtain a better estimate of actual values, the so-called prediction interval has to be considered.

Such interval must take into account both the uncertainty in the estimate of the conditional mean and the variability in the conditional distribution.

In order to find this interval, the so-called prediction error:

$$Y - \hat{Y} = (\beta_0 + \beta_1 X + \varepsilon) - (b_0 + b_1 X)$$

has to be considered.

The expectation for the prediction error is:

$$\begin{aligned} E(Y - \hat{Y}) &= E[(\beta_0 + \beta_1 X + \varepsilon) - (b_0 + b_1 X)] = \\ &= E(\beta_0) + \beta_1 E(X) + E(\varepsilon) - E(b_0) - E(b_1)E(X) = \\ &= \beta_0 + \beta_1 E(X) + 0 - \beta_0 - \beta_1 E(X) = 0 \end{aligned}$$

The **variance of prediction error** can be obtained from the following variances:

$$V(\hat{Y}) = \sigma^2 \left( \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}} \right) \quad V(Y) = V(\beta_0 + \beta_1 X + \varepsilon) = V(\varepsilon)$$

Considering that  $\beta_0$  and  $\beta_1$  are constant values, thus their variance is zero, the following equation can be written (notably,  $V(X) = 0$  in this case, since  $X$  is fixed):

$$\begin{aligned} V(Y - \hat{Y}) &= V[(\beta_0 + \beta_1 X + \varepsilon) - \hat{Y}] = V(\beta_0 + \beta_1 X + \varepsilon) + V(\hat{Y}) = V(\varepsilon) + V(\hat{Y}) = \\ &= \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}} \right) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}} \right) \end{aligned}$$

Given the **normal distribution of Y data**, the following relation can be written:

$$\frac{Y - \hat{Y}}{[V(Y - \hat{Y})]^{1/2}} \sim t_{n-2}$$



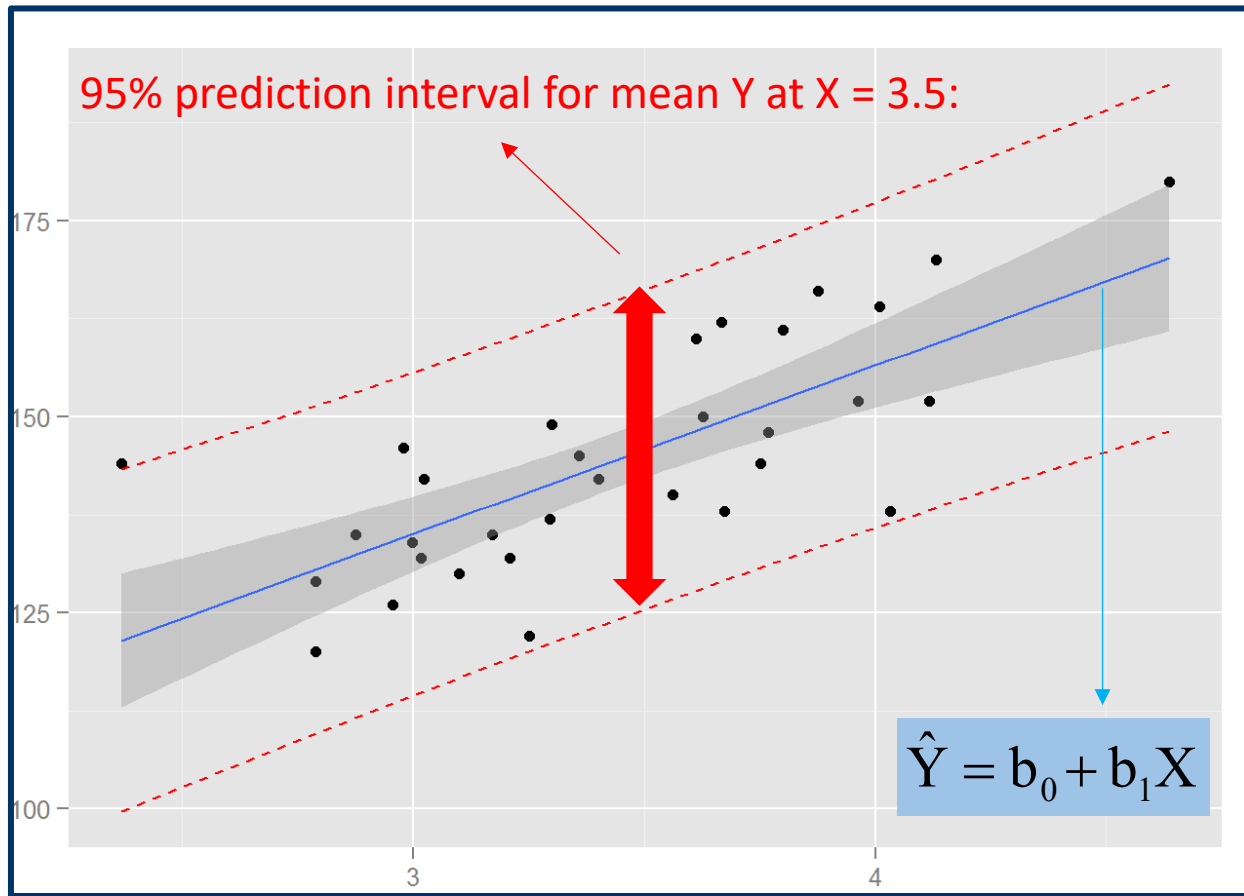
Considering that  $\sigma$  can be estimated by the residual standard deviation,  $s_{y/x}$ , the **prediction interval for Y at a  $\alpha$  level of significance**, can be expressed as:

$$(b_0 + b_1 X) \pm t_{n-2, 1-\alpha/2} S_{Y/X} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}$$

If  $m$  replicates are obtained for response  $Y$  at a specific value of  $X$ , an average value of  $Y$  is obtained, and the corresponding variance becomes  $\sigma^2/m$ , thus the prediction interval for  $Y$  becomes:

$$(b_0 + b_1 X) \pm t_{n+m-3, \alpha/2} S_{Y/X} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}$$

It is apparent that, as emphasized by the figure in the next slide, **the width of the prediction interval is larger than that of the confidence interval** (since the term  $1$ , or  $1/m$ , is present additionally under the square root sign).

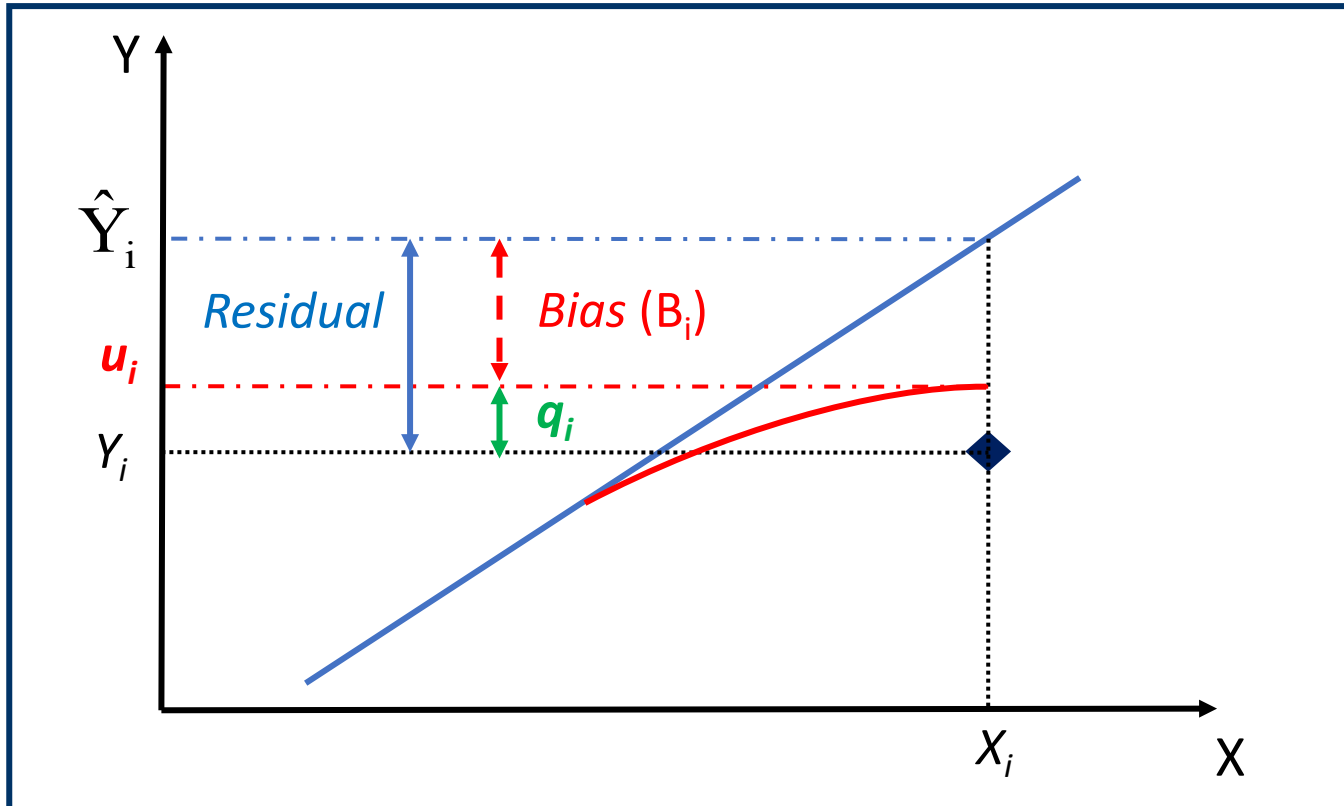


As shown in the figure, if the precision of responses is poor a remarkable difference is observed between prediction and confidence bands.

The two types of bands get closer when the precision of responses is increased.

## Lack of fit in simple linear regression

A key aspect of linear regression is the **evaluation of the eventual lack of fit**, indicating that **the model is inadequate**; it is based on the **analysis of residuals**:



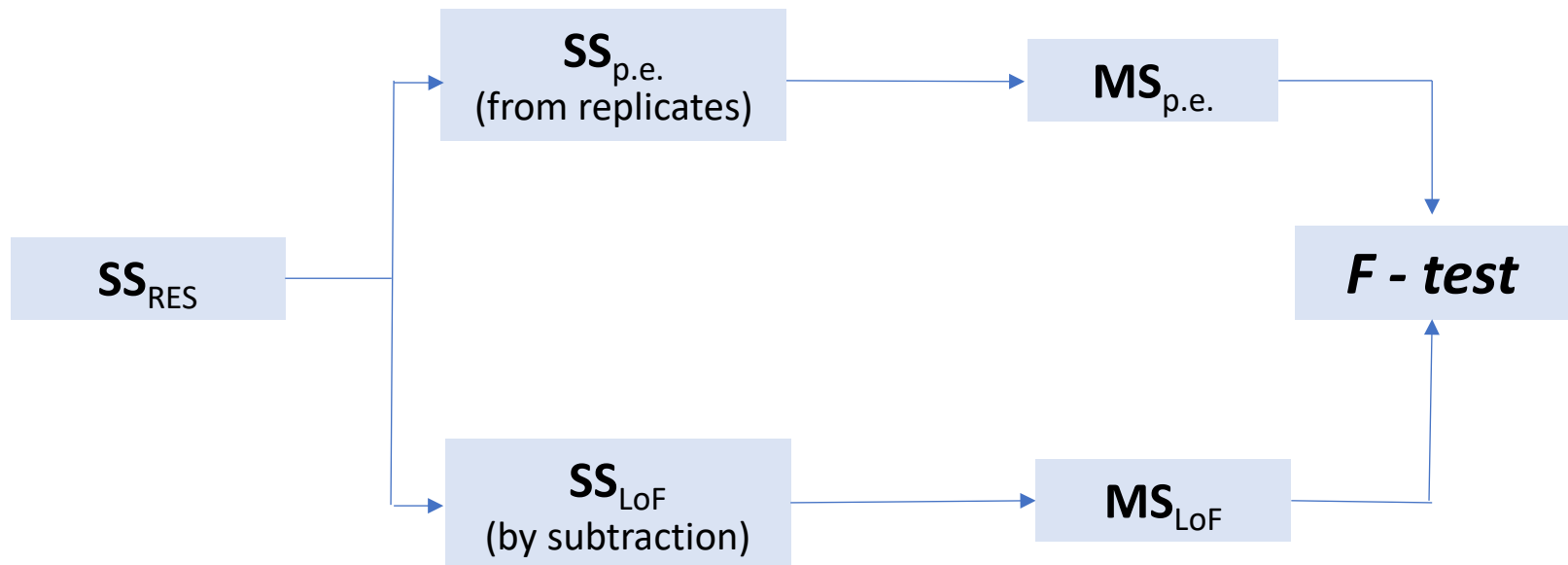
As shown in the figure, if the linear model is not correct (because the curved model is actually correct) a residual includes both a random component,  $q_i$ , and a systematic component, or bias,  $B_i$ , with the latter arising from the inadequacy (lack of fit) of the model.

If a lack of fit exists, the residual mean squares,  $MS_{RES}$ :

$$MS_{RES} = s_{y/x}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}$$

cannot be considered an unbiased estimator of  $\sigma^2$  (the pure error).

Application of ANOVA to residuals analysis enables a separation between contributions due to pure error (p.e.) and to lack of fit (LoF).



An independent estimate of  $\sigma^2$ , easily obtained by replicating responses at each  $X_i$ , is required:

$$Y_{11}, Y_{12}, \dots, Y_{1j}, \dots, Y_{1n_1} \Rightarrow \bar{Y}_1; s_1^2 \quad n_1 \text{ replicates at } X_1$$

$$Y_{21}, Y_{22}, \dots, Y_{2j}, \dots, Y_{2n_2} \Rightarrow \bar{Y}_2; s_2^2 \quad n_2 \text{ replicates at } X_2$$

$$Y_{i1}, Y_{i2}, \dots, Y_{ij}, \dots, Y_{in_i} \Rightarrow \bar{Y}_i; s_i^2 \quad n_i \text{ replicates at } X_i$$

$$Y_{h1}, Y_{h2}, \dots, Y_{hj}, \dots, Y_{hn_h} \Rightarrow \bar{Y}_h; s_h^2 \quad n_h \text{ replicates at } X_h$$

Considering  $n_1 = n_2 = \dots = n_h = n$ , the pure error sum of squares,  $SS_{p.e.}$ , is:


$$SS_{p.e.} = \sum_{i=1}^h \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$$

Since the number of degrees of freedom is given by  $N - h$ , where:  $N = \sum_{i=1}^h n_i$


the pure error mean square,  $MS_{p.e.}$ , is calculated as:

$$MS_{p.e.} = \frac{\sum_{i=1}^h \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2}{N - h}$$

The ANOVA table for simple linear regression, also including the contribution due to the Lack of fit, is the following:



Source of variation	SS	df	MS	E(MS)
Total	$\sum_{i=1}^h \sum_{j=1}^n (Y_{ij} - \bar{Y})^2$	$N-1$		
Regression	By subtraction	1	$MS_{REG}$	$\sigma^2 + \beta_1^2 S_{XX}$
Residuals	$\sum_{i=1}^h \sum_{j=1}^n (Y_{ij} - \hat{Y}_i)^2$	$N-2$	$MS_{RES}$	$\sigma^2$
Lack of Fit	By subtraction	$h-2$	$MS_{LoF}$	$\sigma_{p.e.}^2 + \frac{\sum_i B_i^2}{N-2}$
Pure error	$\sum_{i=1}^h \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$	$N-h$	$MS_{p.e.}$	$\sigma_{p.e.}^2$



In order to find the presence of lack of fit the following statistic needs to be defined:

$$F = MS_{LoF}/MS_{p.e}$$

which is distributed according to an F distribution with h-2, N-h degrees of freedom.

The presence of Lack of fit will thus be confirmed, with a  $\alpha$  significance coefficient, if:

$$MS_{LoF}/MS_{p.e} \geq F_{h-2, N-h, \alpha}$$

It can be demonstrated that, as reported in the ANOVA table shown before, the expected value for  $MS_{LoF}$  is:

$$\sigma_{p.e.}^2 + \frac{\sum_i B_i^2}{N-2}$$

Since  $MS_{p.e.} = \sigma_{p.e.}^2$ , the absence of a significant difference between  $MS_{LoF}$  and  $MS_{p.e.}$  implies that all the bias values  $B_i$  are negligible (indeed, the sum of their squares is close to 0), which is consistent with the absence of lack of fit.

## A numerical example

Suppose that the following X and Y values were obtained from a series of measurements:

X	Y
90	81; 83
79	75
66	68; 60; 62
51	60; 64
35	51; 53

These sums can be easily calculated:

$$\sum_{i=1}^{10} x_i = 629, \sum_{i=1}^{10} y_i = 657, \sum_{i=1}^{10} x_i^2 = 43161, \sum_{i=1}^{10} y_i^2 = 44249, \sum_{i=1}^{10} x_i y_i = 43189$$

The **total sum of squares**,  $SS_{\text{TOT}}$ , is thus given by:

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 44249 - 10(65.7)^2 = 1084.1$$



The slope of regression line is:

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{10} x_i y_i - 10 \bar{x} \bar{y}}{\sum_{i=1}^{10} x_i^2 - 10 \bar{x}^2} = \frac{43189 - 10 * 62.9 * 65.7}{43161 - 10 * (62.9)^2} = 0.51814$$

The regression sum of squares,  $SS_{\text{REG}}$ , is:

$$b_1^2 S_{XX} = 965.65636$$

As a result, the residual sum of squares is  $1084.1 - 965.66 = 118.44$ .

Calculations required for the pure error sum of squares,  $SS_{p.e.}$ , are:

$X_i$   
↓

90:  $\bar{Y}_1 = \frac{81+83}{2} = 82 \quad \longrightarrow \quad \sum_{i=1}^2 (Y_{1i} - \bar{Y}_1)^2 = (81-82)^2 + (83-82)^2 = 2$

79:  $\bar{Y}_2 = 75 \quad \longrightarrow \quad \sum_{i=1}^2 (Y_{2i} - \bar{Y}_1)^2 = (75-75)^2 = 0$

66:  $\bar{Y}_3 = \frac{68+60+62}{3} = 63.33 \quad \longrightarrow \quad \sum_{i=1}^3 (Y_{3i} - \bar{Y}_3)^2 = (68-63.33)^2 + (60-63.33)^2 + (62-63.33)^2 = 34.67$

51:  $\bar{Y}_4 = \frac{60+64}{2} = 62 \quad \longrightarrow \quad \sum_{i=1}^2 (Y_{4i} - \bar{Y}_4)^2 = (60-62)^2 + (64-62)^2 = 8$

35:  $\bar{Y}_5 = \frac{51+53}{2} = 52 \quad \longrightarrow \quad \sum_{i=1}^2 (Y_{5i} - \bar{Y}_5)^2 = (51-52)^2 + (53-52)^2 = 2$

$$SS_{p.e.} = 2 + 0 + 34.67 + 8 + 2 = 46.67$$

$$SS_{LoF} = SS_{RES} - SS_{p.e.} = 118.44 - 46.67 = 71.77$$

The following ANOVA table is thus obtained:

<i>Source of variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>
<b><i>Regression</i></b>	965.66	1	965.66
<b><i>Lack of Fit</i></b>	71.77	3	23.92
<b><i>Pure error</i></b>	46.67	5	9.33
<b><i>Total</i></b>	1084.1	9	

Since the ratio  $F = MS_{\text{LoF}}/MS_{\text{p.e.}}$  is lower than the critical value of F distribution:

$$F = \frac{23.92}{9.33} = 2.56 < f_{3,5,0.05} = 5.41$$

The simple linear model is adequate, at a level of confidence of 95%.

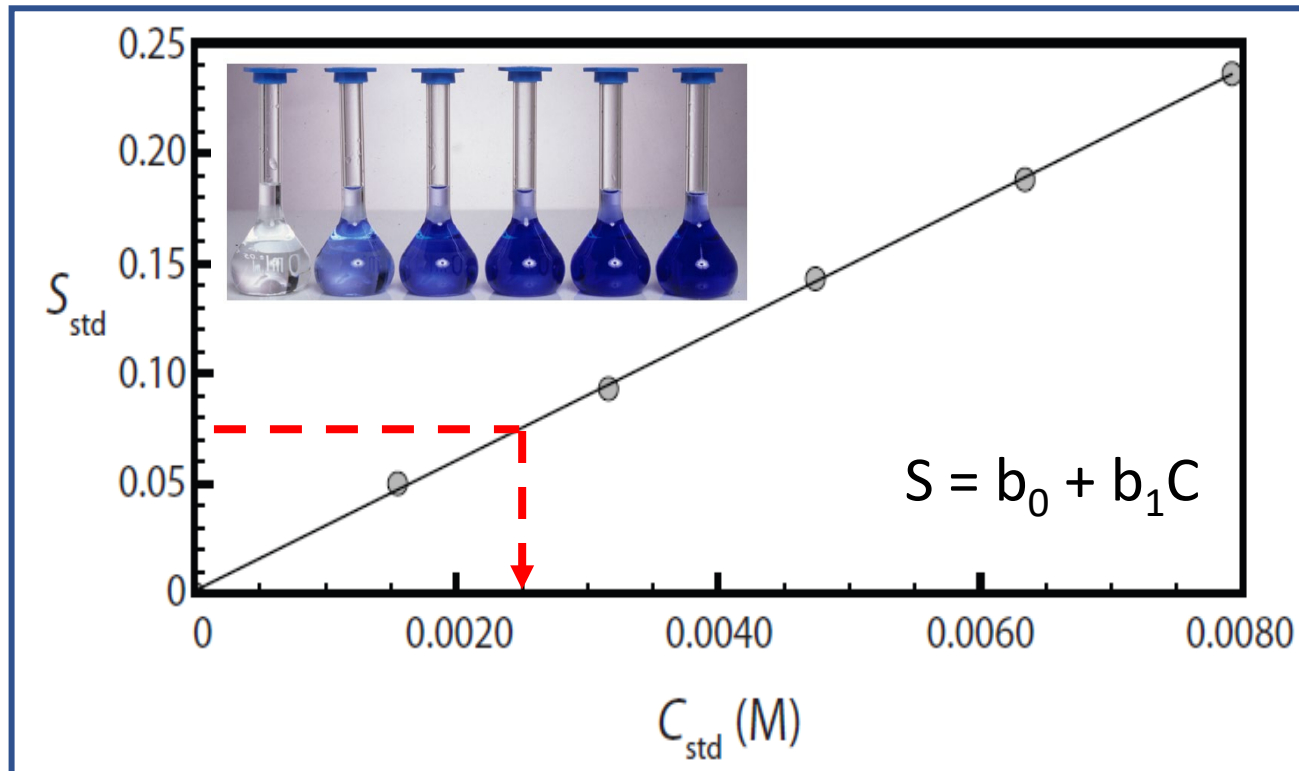
# Calibration and inverse regression

A calibration experiment typically consists of **two stages**.

In the first stage  $n$  observations  $(x_i, y_i)$  are collected from standards and used to fit a **regression model**, that in the simplest case can be expressed as:

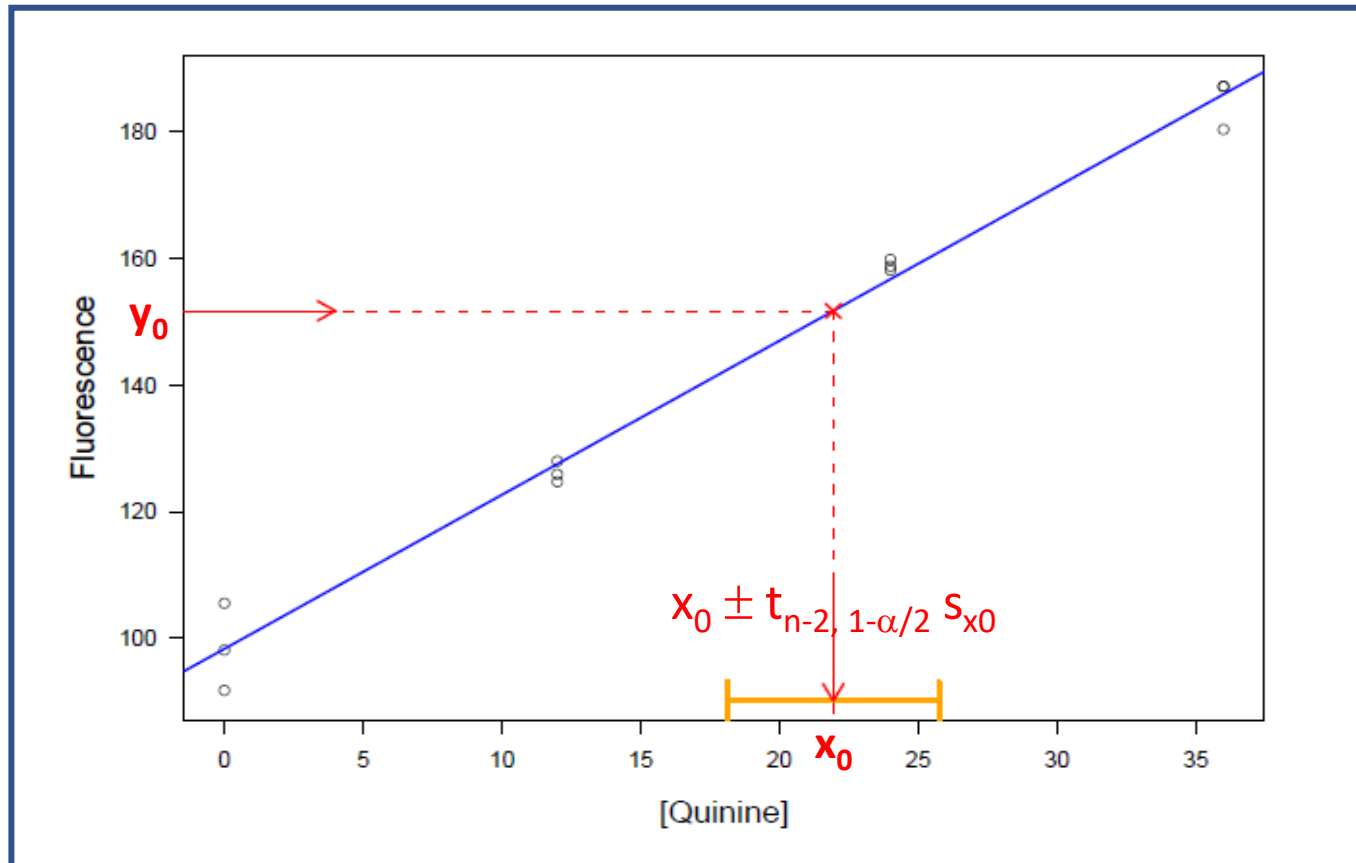
$$y = \beta_0 + \beta_1 x + \varepsilon$$

The fitted model is often referred to as the **calibration curve or standards curve**.



In the second stage  $m$  ( $m \geq 1$ ) values of the response are observed for an unknown predictor value  $x_0$ .

Inverse regression (also called inverse prediction or discrimination) is the procedure, based on the calibration line, adopted to estimate the value  $x_0$  from a single observation  $y_0$  or from the mean of  $m$  replicated observations.



The  $x_0$  estimate is based on inverting the calibration curve at  $y_0$  (i.e., solving the fitted regression equation for  $x_0$ ) and is easily extended to polynomial and nonlinear calibration problems:

single observation)  $\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} \Leftrightarrow \hat{x}_0 = \bar{x} + \frac{(y_0 - \bar{y})}{\hat{\beta}_1} = \bar{x} + \frac{S_{xx}}{S_{xy}}(y_0 - \bar{y})$

m replicates)  $\hat{x}_0 = \frac{\bar{y}_{0m} - \hat{\beta}_0}{\hat{\beta}_1}$  where:  $\bar{y}_{0m} = \frac{1}{m} \sum_{i=1}^m y_{0i}$

Two approaches are usually adopted to calculate confidence intervals for  $\hat{x}_0$ :

- 1) Wald interval
- 2) Inversion interval

## Wald interval

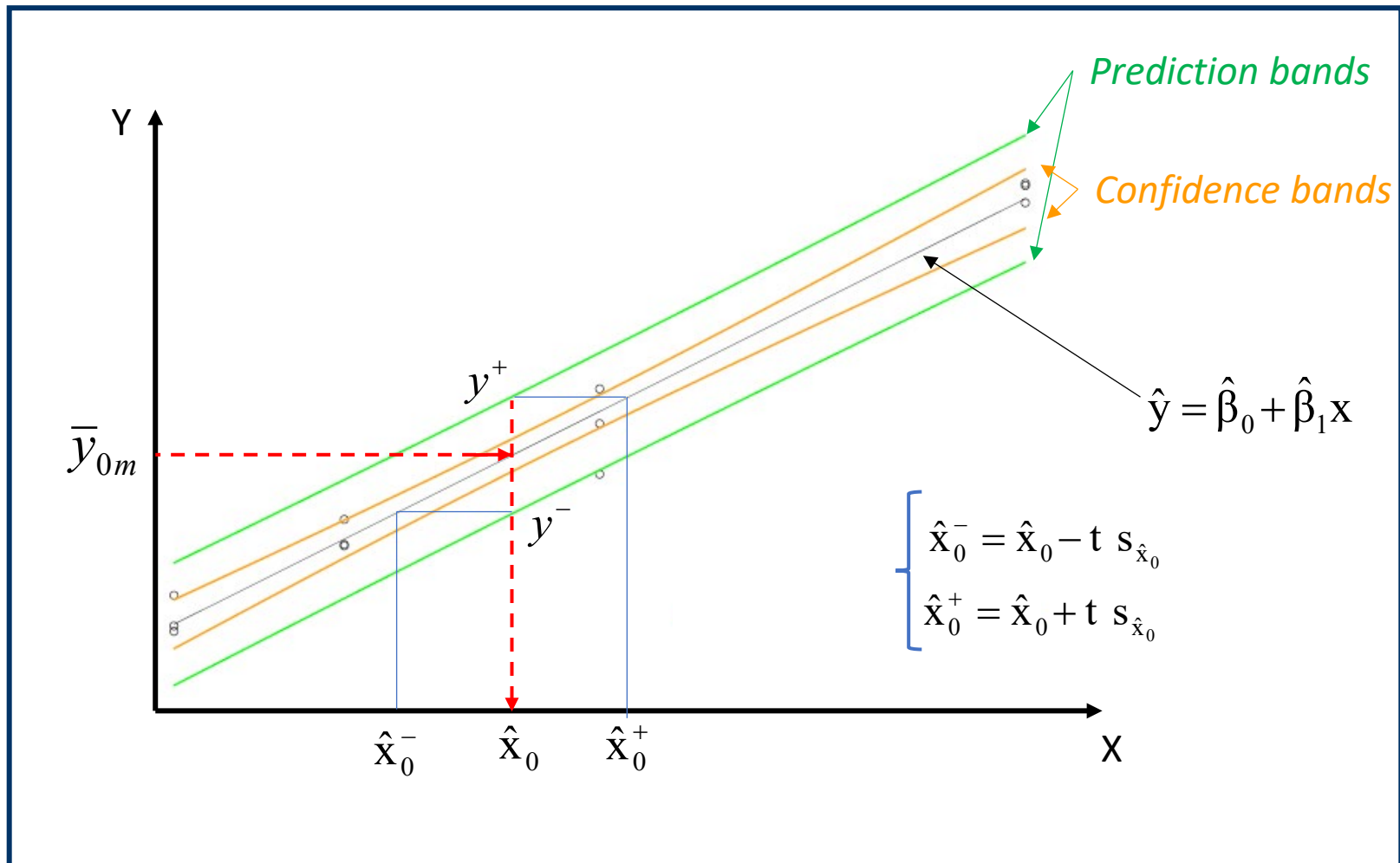
According to the Wald approach, the following standard deviation can be estimated for  $\hat{x}_0$  :

$$s_{\hat{x}_0} = \frac{s_{y/x}}{\hat{\beta}_1} \left[ \frac{1}{m} + \frac{1}{n} + \frac{(\hat{x}_0 - \bar{x})^2}{S_{xx}} \right]^{1/2}$$

thus an approximate 100 (1- $\alpha$ ) % confidence interval for  $\hat{x}_0$  is given by:

$$\hat{x}_0 \pm t_{(1-\alpha/2), n+m-3} s_{\hat{x}_0} = \hat{x}_0 \pm t_{(1-\alpha/2), n+m-3} \frac{s_{y/x}}{\hat{\beta}_1} \left[ \frac{1}{m} + \frac{1}{n} + \frac{(\hat{x}_0 - \bar{x})^2}{S_{xx}} \right]^{1/2}$$

From a graphical point of view this interval can be obtained by exploiting prediction bands, namely, by extrapolating x values corresponding to the upper ( $y^+$ ) and lower ( $y^-$ ) limits of the prediction interval for  $y_0$ :





For example, if the  $y^+$  value corresponding to  $x = x_0$  is considered:

$$y^+ = (b_0 + b_1 \hat{x}_0) + t_{n+m-3, \alpha/2} s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}$$

$x_0^+$  corresponds to the abscissa of the calibration line point that has a  $y^+$  ordinate:

$$y^+ = (b_0 + b_1 \hat{x}_0) + t_{n+m-3, \alpha/2} s_{y/x} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}} = (b_0 + b_1 \hat{x}_0^+)$$

thus:

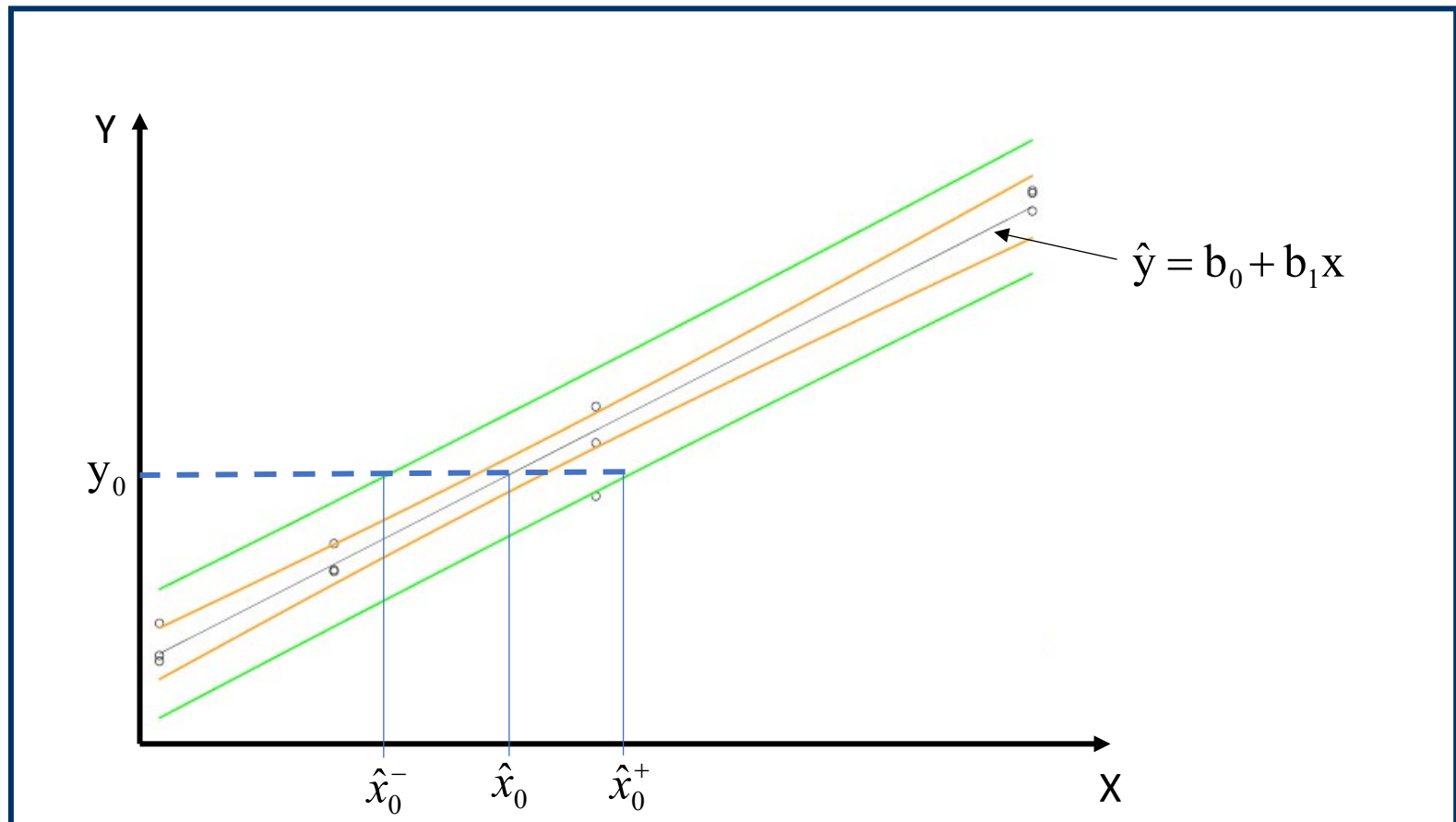
$$\hat{x}_0^+ = \hat{x}_0 + t_{n+m-3, \alpha/2} \frac{s_{y/x}}{b_1} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}$$

which means that  $x_0^+$  is the upper end of the confidence interval calculated for  $x_0$  using the Ward equation.

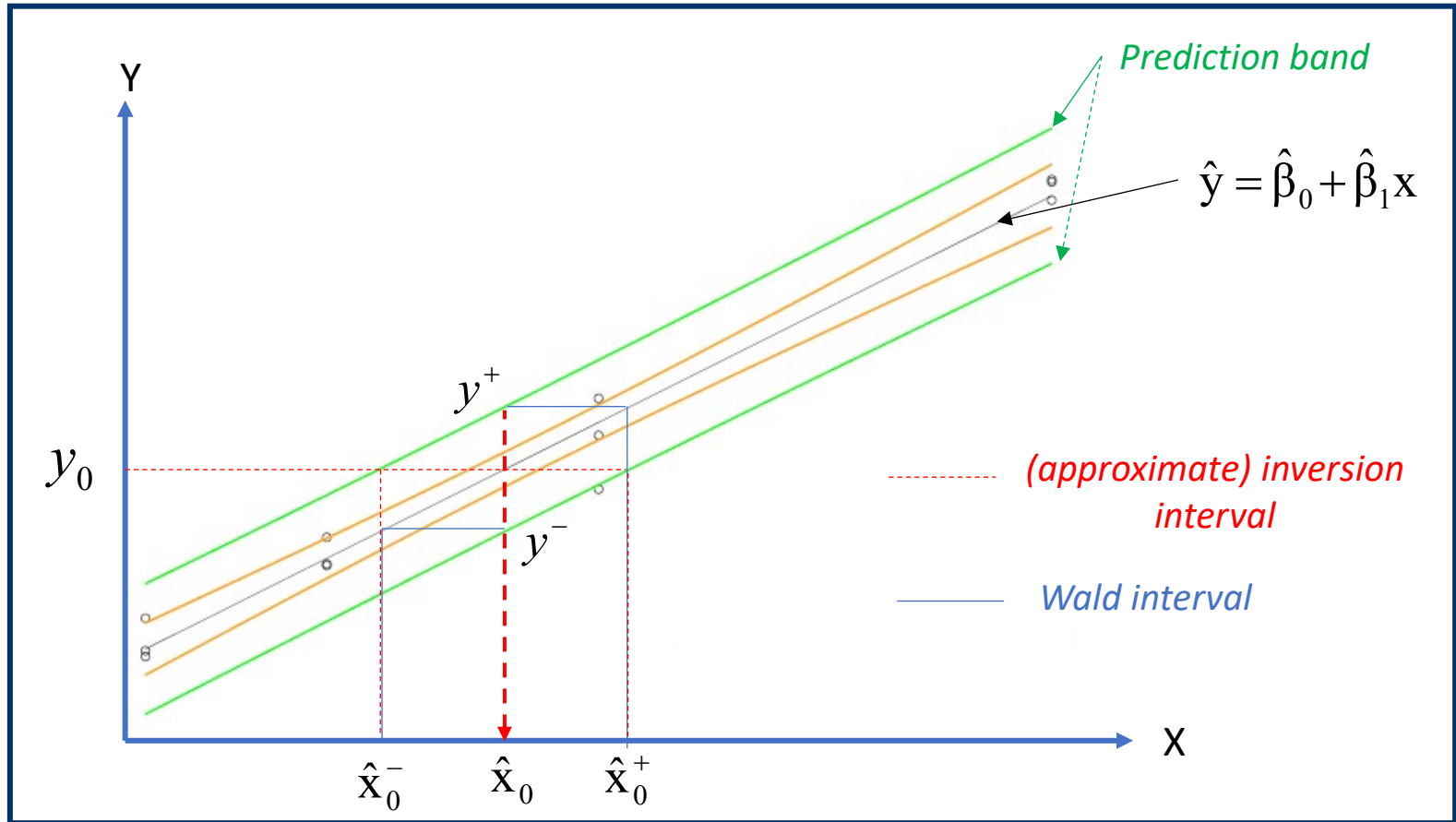
## Inversion interval

A confidence interval for  $\hat{x}_0$  can be constructed also by inverting a prediction interval for the response, thus it is called “inversion interval”.

From a graphical point of view, an approximate inversion interval can be obtained by drawing a horizontal line through the scatterplot of the standards at  $y_0$  and finding the abscissas of its intersections with the prediction bands of the calibration curve:



Interestingly, for the simple linear calibration problem the Wald-based interval is equivalent to the approximate inversion interval:



It is worth noting that the expression for  $s_{\hat{x}_0}$  can be expressed in an alternative form, if the following relationships are considered:

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \quad \longrightarrow \quad \hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

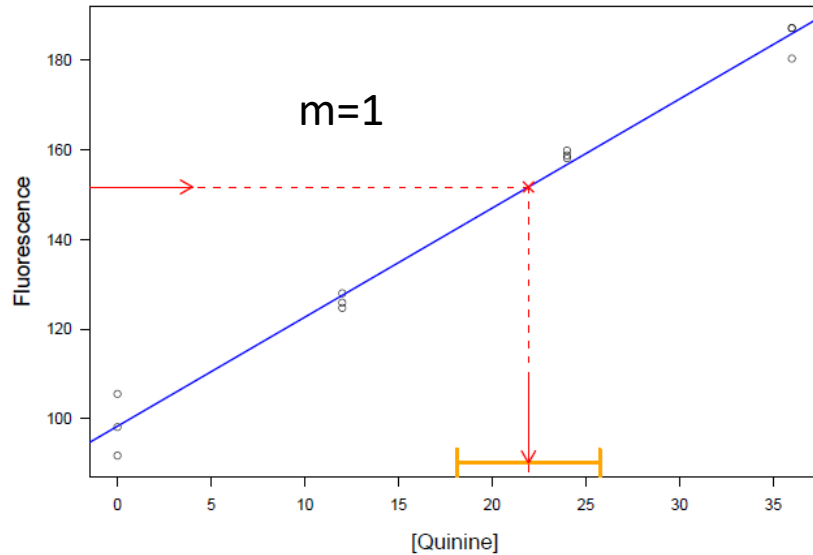
Indeed, the equation:

$$s_{\hat{x}_0} = \frac{s_{y/x}}{\hat{\beta}_1} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

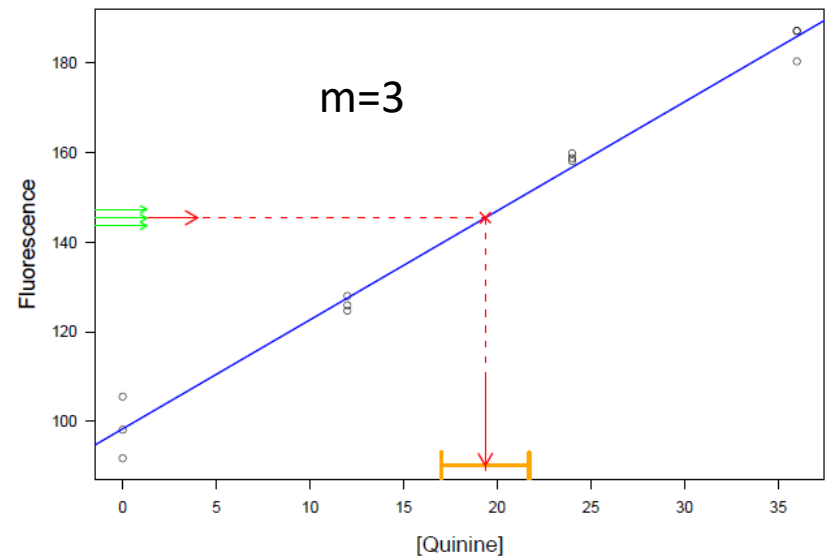
is easily transformed into:

$$s_{\hat{x}_0} = \frac{s_{y/x}}{\hat{\beta}_1} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_0 - \bar{y})^2}{\hat{\beta}_1^2 S_{xx}}}$$

As shown in the following figure, even a limited number of replicates can reduce significantly the uncertainty on  $x_0$ :



$$s_{\hat{x}_0} = \frac{s_{y/x}}{\hat{\beta}_1} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$



$$s_{\hat{x}_0} = \frac{s_{y/x}}{\hat{\beta}_1} \sqrt{\frac{1}{3} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Moreover, given a certain number of replicates, the uncertainty on  $x_0$  is minimized when  $x_0 = \bar{x}$  and when  $S_{xx}$  is increased.

## Considerations about the $S_{xx}$ term and the experimental design of calibration

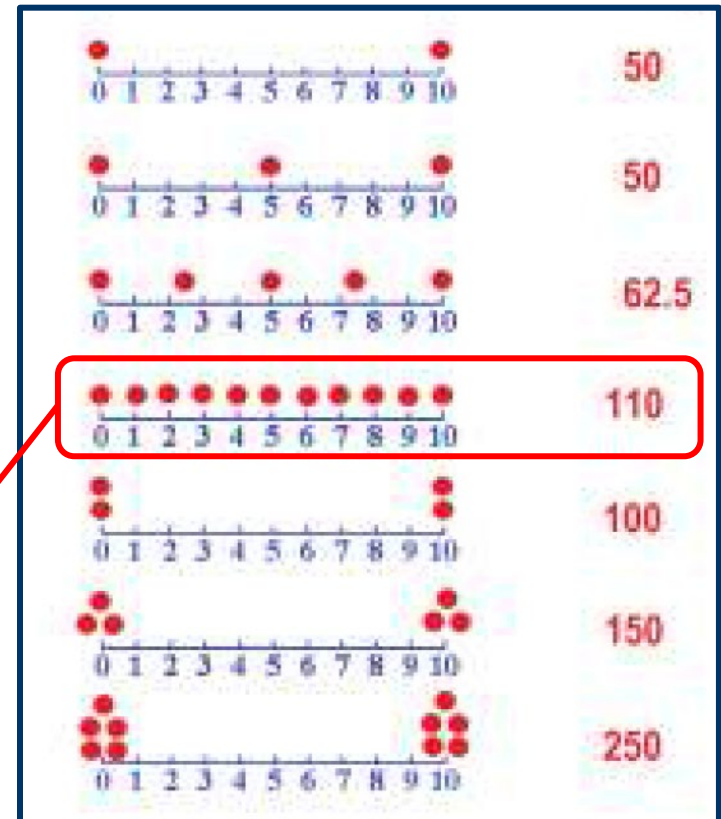
The  $S_{xx}$  term depends on the number of experimental points but also on their distribution along the x axis.

$S_{xx}$  values as a function of the distribution of calibration points are reported in the figure on the right:

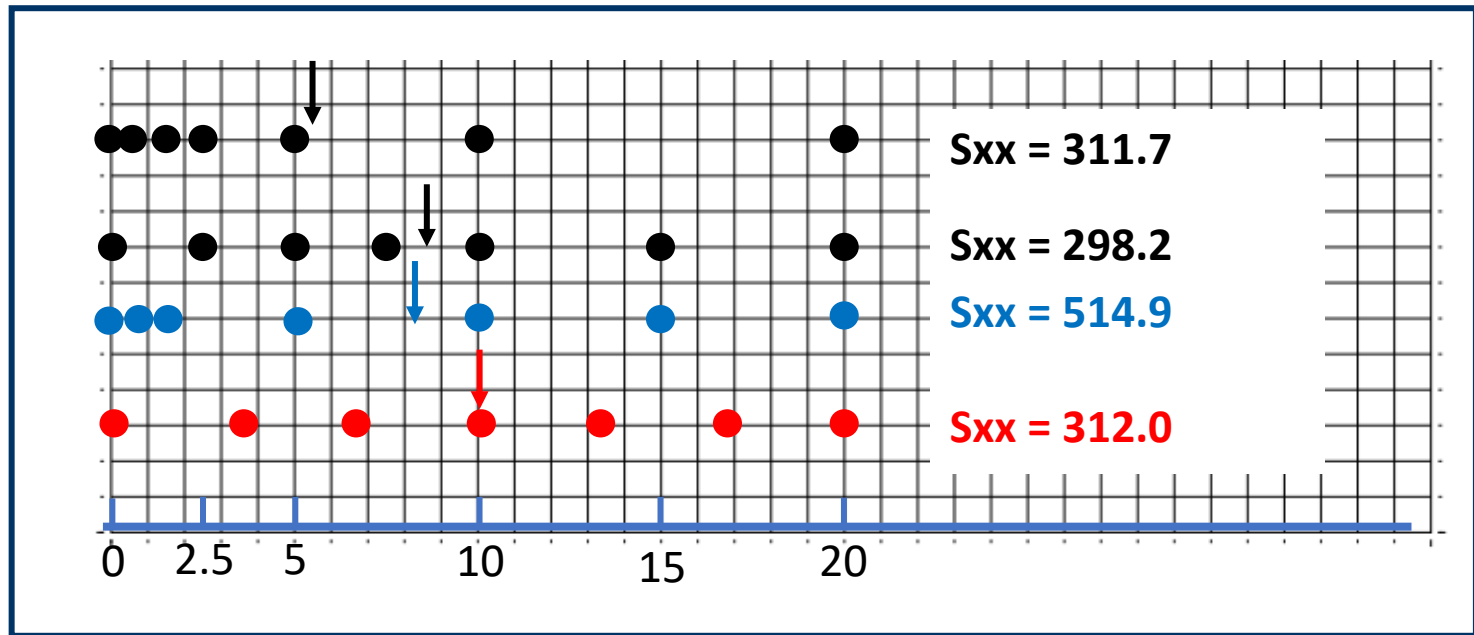
As apparent, they are higher for designs including clusters of points close to the ends of the x interval.

However, such designs would not provide good guarantees on the adequacy of the adopted linear regression model.

Consequently, calibration designs including evenly-spaced points, that lead to intermediate values for  $S_{xx}$  but, additionally, provide information on important aspects like homoscedasticity and linearity, are usually preferred.



It is also worth noting that, as shown in the following figure, the distribution of experimental points has a clear influence also on the centroid, indicated by an arrow in each set of points:



The centroid displacement has, in turn, a direct influence on the position of confidence and prediction bands, and, consequently, on the width of confidence intervals for extrapolated concentrations.

The choice of the calibration design has thus to consider also this important aspect.