# Linear Regression and Orthogonal Regression in Space and Spacetime

Martin Erik Horn (Email: m.horn@iubh.de)

OHP Slides for the tutorials
“Advanced Statistics – DLMDSAS01”
and “Advanced Mathematics – DLMDSAM01”
of the
***Online International*** Master program
“Computer Science”
at
**iu** – Internationale Hochschule /
International University of Applied Sciences,
Campus Berlin, Summer 2021

(Version: July 6th, 2021)

# Overview / Tutorial Dates:

Mon, June 14th, 2021 (CS virtual 928):
Advanced Statistics

$\rightarrow$ Linear Regression

Mon, June 28th, 2021 (CS virtual 928):
Advanced Statistics

$\rightarrow$ Orthogonal Regression in Space

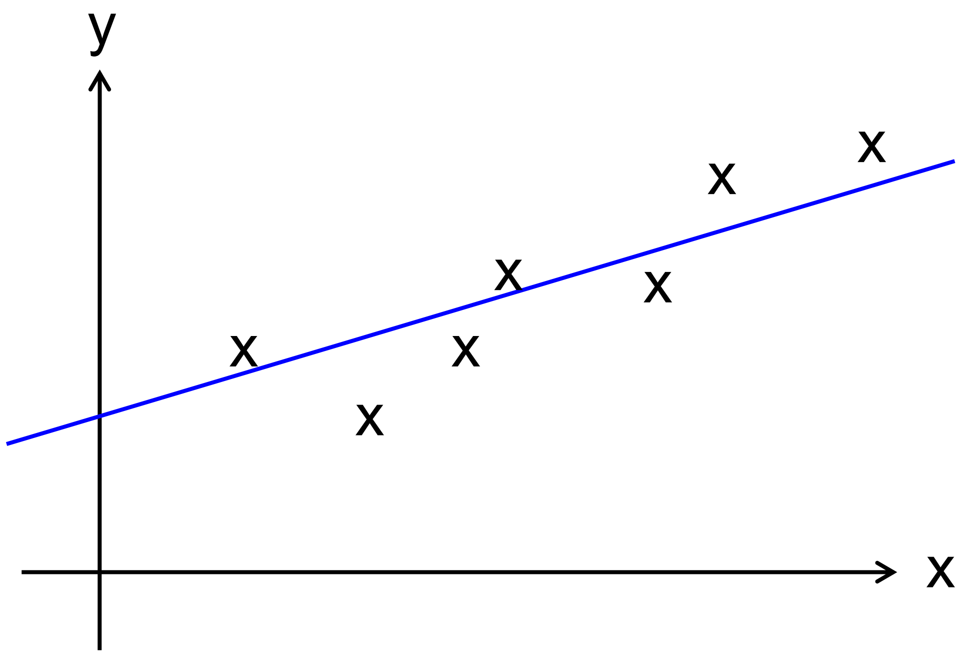
Mon, July 5th, 2021 (CS virtual 928):
Advanced Mathematics

$\rightarrow$ Orthogonal Regression in Spacetime

# Regression Analysis

*(Regressionsanalyse)*

Problem: If we assume that there is a linear relation between the values of two variables, how can we find the line of best fit?

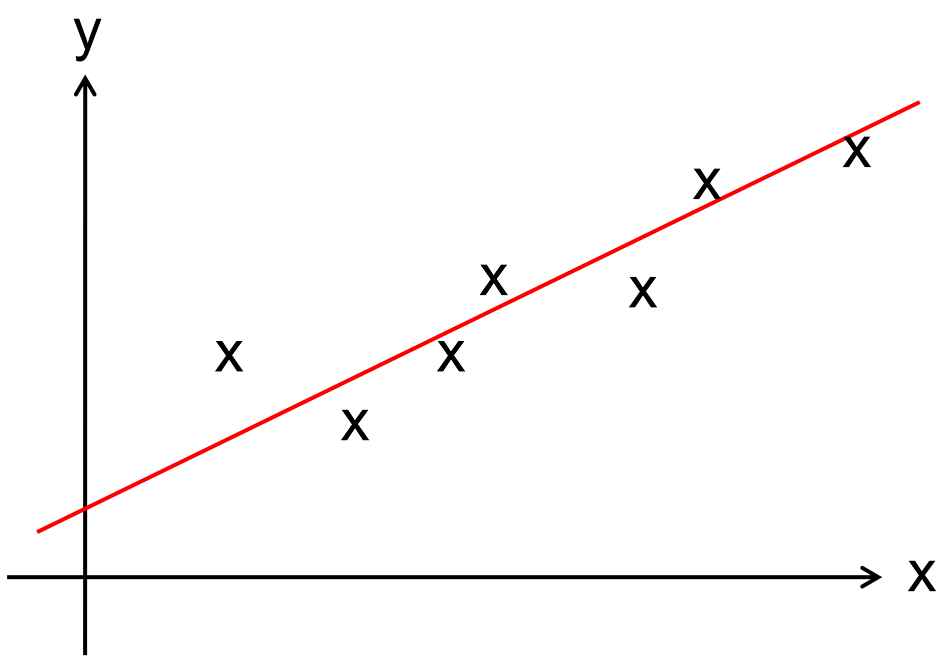



First possible line of best fit

# <u>Regression Analysis</u>

*(Regressionsanalyse)*

Problem: If we assume that there is a linear relation between the values of two variables, how can we find the line of best fit?

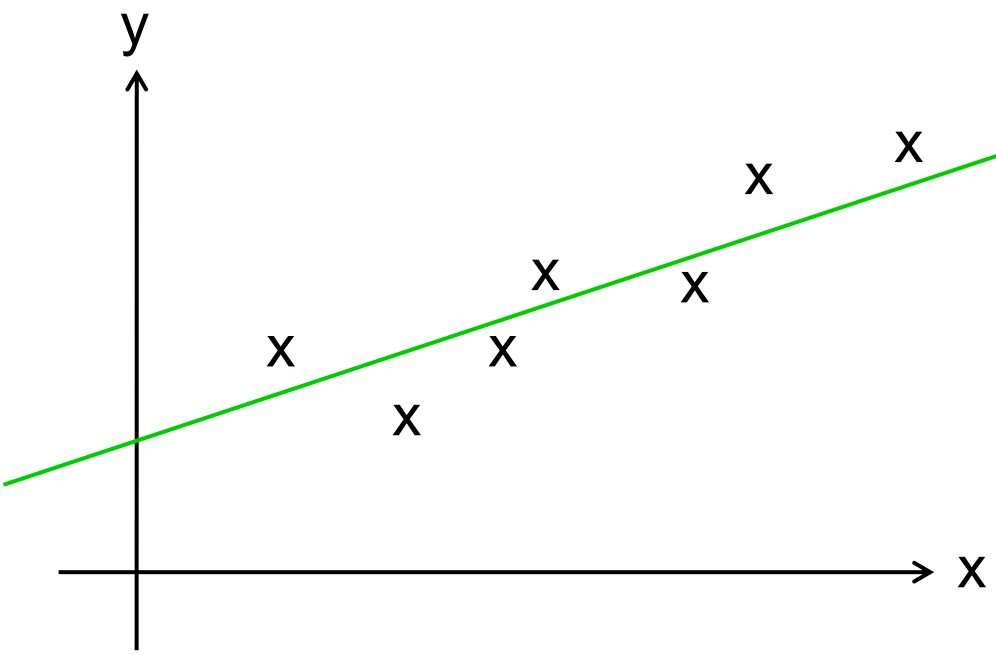



Second possible line of best fit

# Regression Analysis

*(Regressionsanalyse)*

Problem: If we assume that there is a linear relation between the values of two variables, how can we find the line of best fit?

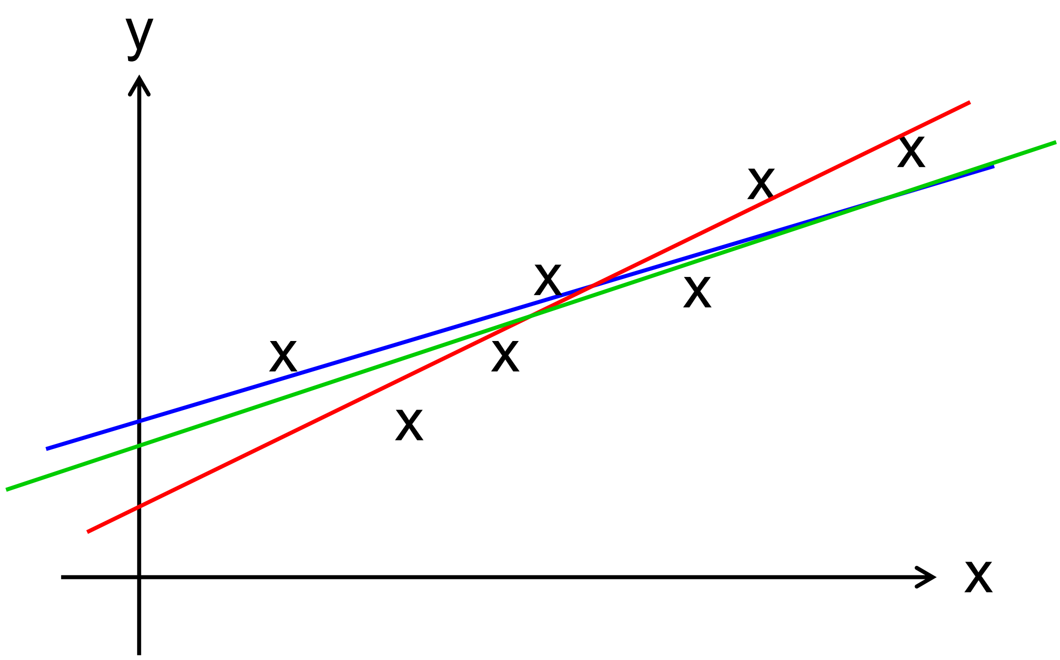



Third possible line of best fit

# Regression Analysis

*(Regressionsanalyse)*

Problem: If we assume that there is a linear relation between the values of two variables, how can we find the line of best fit?




What is the line of best fit?

⇒ We will find the line of best fit by finding the least-square line.

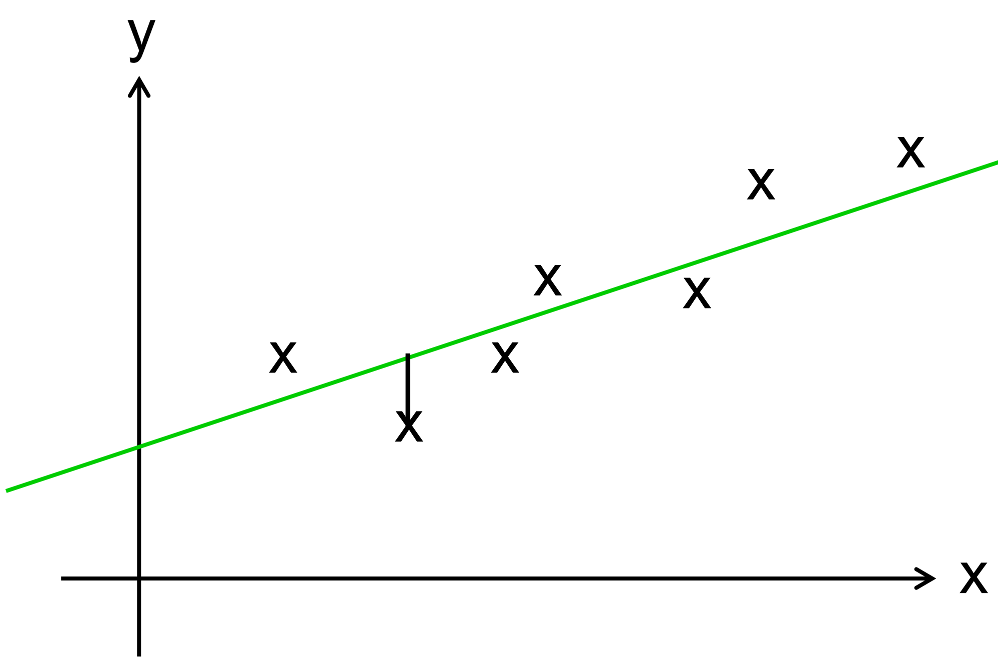
# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

The least-square line (regression line) will be written as:

$$y = m\,x + b$$

slope of the regression line ($m$) — y-intercept of the regression line ($b$)



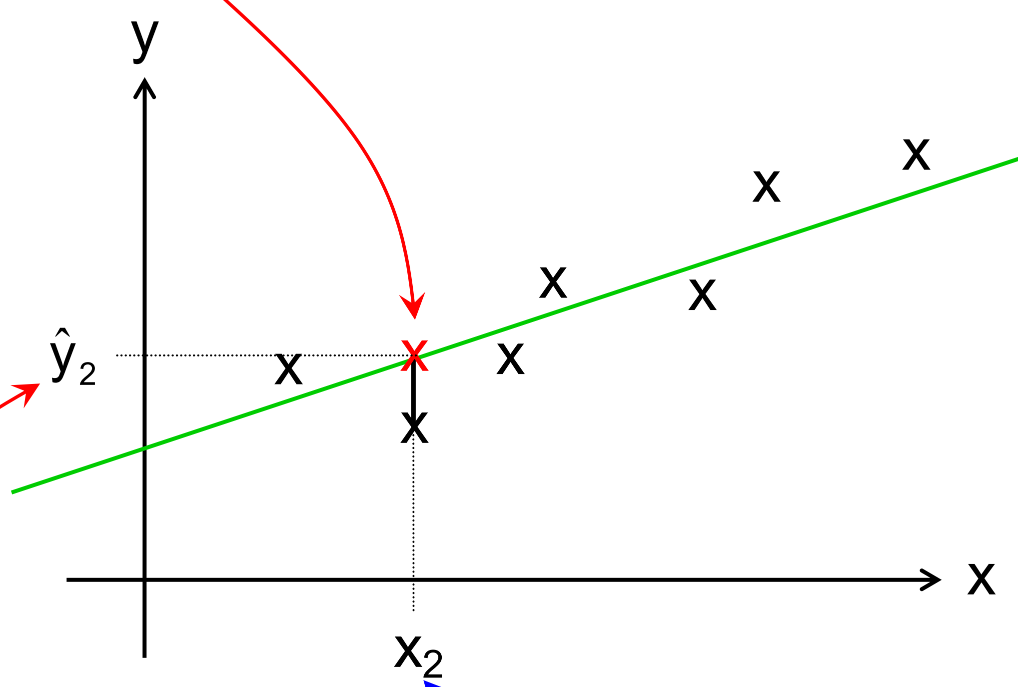

Actual values of the data points:

$(x_1; y_1)$ $(x_2; y_2)$ $(x_3; y_3)$ $(x_4; y_4)$ …

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

The estimated values of y are called ŷ (y hat). They are the values of y which lie on the regression line.




Actual values of the data points:

$(x_1; y_1)$ $(x_2; y_2)$ $(x_3; y_3)$ $(x_4; y_4)$ …

Estimated values of the data points:

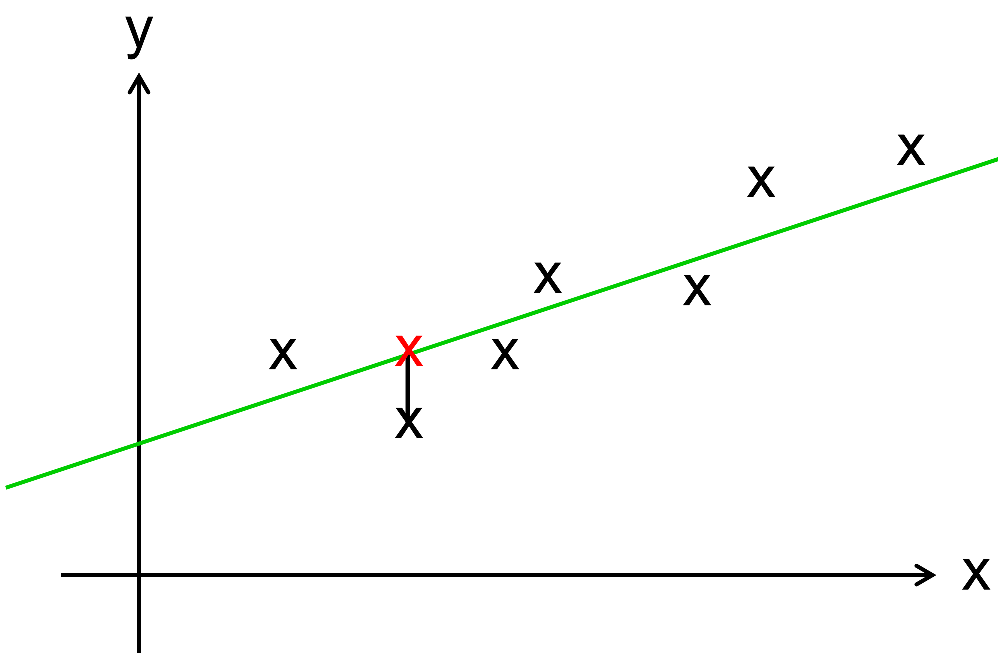
$(x_2; \hat{y}_2)$ …

identical value of $x_2$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

The estimated values of y are called $\hat{y}$ (y hat). They are the values of y which lie on the regression line.




Actual values of the data points:

$(x_1;\ y_1)$ $(x_2;\ y_2)$ $(x_3;\ y_3)$ $(x_4;\ y_4)$ . . .

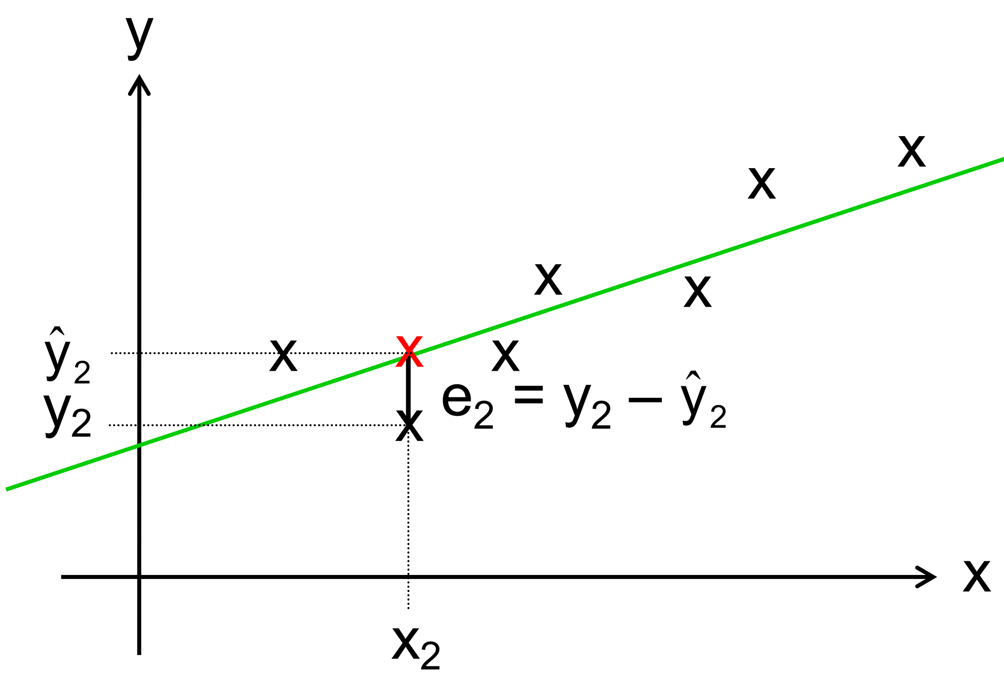
Estimated values of the data points:

$(x_1;\ \hat{y}_1)$ $(x_2;\ \hat{y}_2)$ $(x_3;\ \hat{y}_3)$ $(x_4;\ \hat{y}_4)$ . . .

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

The distance between the actual value and the estimated value of y is called error or residual.




Actual values of the data points:

$(x_1; y_1)$ $(x_2; y_2)$ $(x_3; y_3)$ $(x_4; y_4)$ …

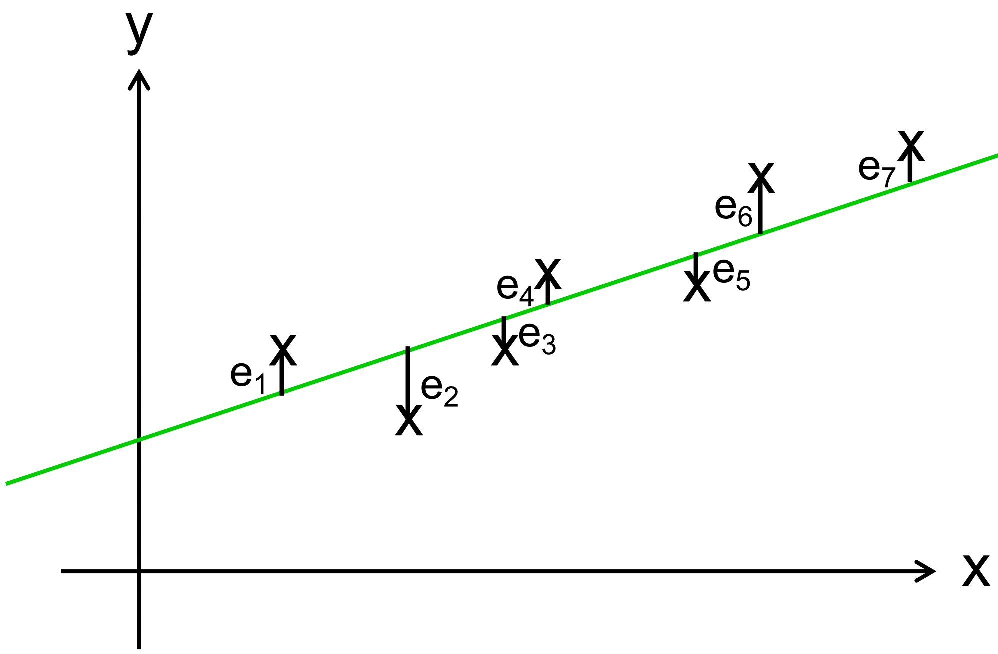
Estimated values of the data points:

$(x_1; \hat{y}_1)$ $(x_2; \hat{y}_2)$ $(x_3; \hat{y}_3)$ $(x_4; \hat{y}_4)$ …

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

The distance between the actual value and the estimated value of y is called error or residual.




Residuals (errors) of the data points:

$$e_1 = y_1 - \hat{y}_1 = y_1 - m\,x_1 - b$$

$$e_2 = y_2 - \hat{y}_2 = y_2 - m\,x_2 - b$$

$$e_3 = y_3 - \hat{y}_3 = y_3 - m\,x_3 - b$$

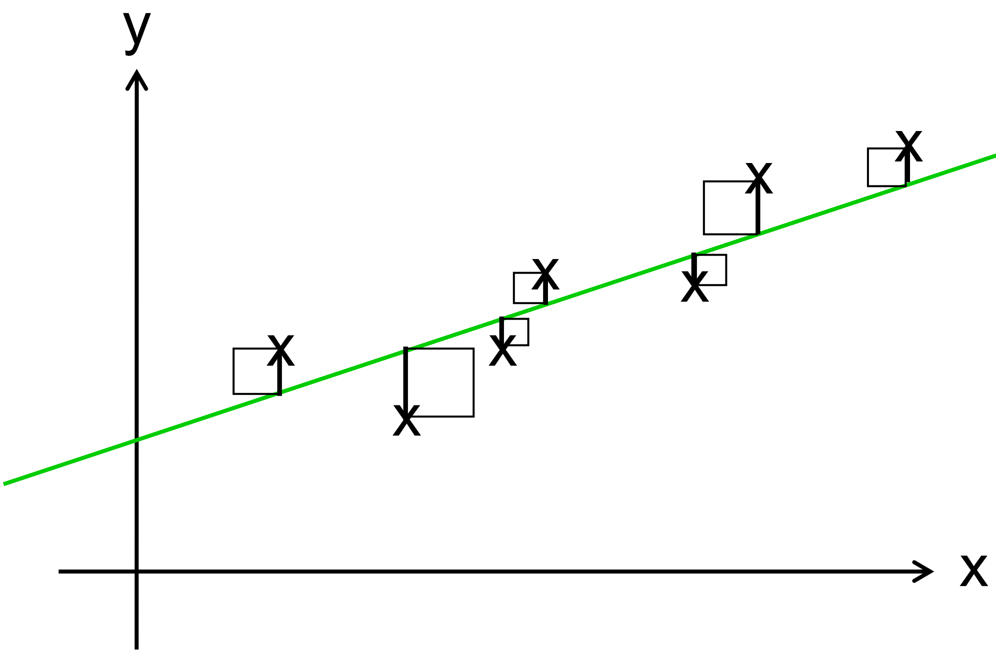
$$e_4 = y_4 - \hat{y}_4 = y_4 - m\,x_4 - b$$

etc …

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Now the errors or residuals are squared …




Squared errors (squared residuals):

$$e_1^2 = (y_1 - \hat{y}_1)^2 = (y_1 - m\,x_1 - b)^2$$

$$e_2^2 = (y_2 - \hat{y}_2)^2 = (y_2 - m\,x_2 - b)^2$$

$$e_3^2 = (y_3 - \hat{y}_3)^2 = (y_3 - m\,x_3 - b)^2$$

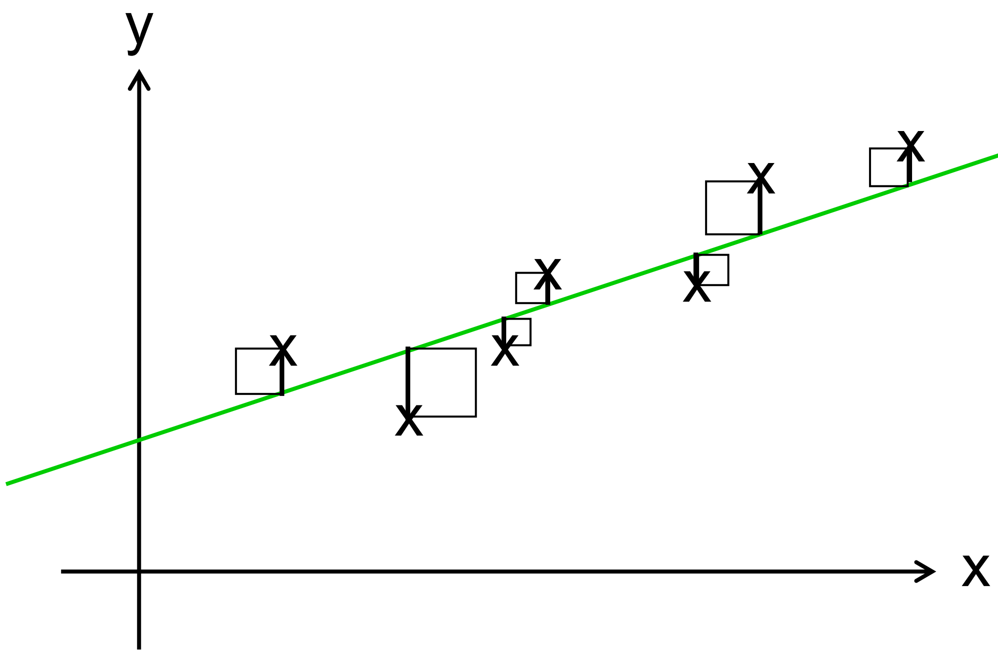
$$e_4^2 = (y_4 - \hat{y}_4)^2 = (y_4 - m\,x_4 - b)^2$$

etc …

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Now the errors or residuals are squared and added.




Sum of squared errors (SSE),
Sum of squared residuals:

$$SSE = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Now the errors or residuals are squared and added and we get the sum of squared errors:

$$SSE = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

The least-square line is the line at which the sum of squared errors has a minimum.

The values of the data points $(x_i; y_i)$ are given because they had been measured.

Only the values of the two parameters m and b can be changed.

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Now the errors or residuals are squared and added and we get the sum of squared errors:

$$SSE = \sum_{i=1}^{n} e_i^{\,2} = e_1^{\,2} + e_2^{\,2} + e_3^{\,2} + \ldots + e_n^{\,2}$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

The least-square line is the line at which the sum of squared errors has a minimum.

The values of the data points $(x_i; y_i)$ are given because they had been measured.

⇒ Thus the sum of squared errors (SSE) depends on the two variables m and b.

⇒ We have to find the minimum of the sum of squared errors (SSE) with respect to these two variables m and b.

⇒ The partial derivatives of the sum of squared errors with respect to m and b are required.

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

Partial derivative of the sum of squared errors with respect to the variable b (y-intercept):

$$\frac{\partial\,\text{SSE}}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} (y_i - m\,x_i - b)^2\right)$$

$$= \sum_{i=1}^{n} \left(\frac{\partial}{\partial b} (y_i - m\,x_i - b)^2\right)$$

$$= \sum_{i=1}^{n} \big(2\,(y_i - m\,x_i - b)\,(-1)\big)$$

$$= -2 \sum_{i=1}^{n} (y_i - m\,x_i - b)$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

Partial derivative of the sum of squared errors with respect to the variable b:

$$\frac{\partial\, SSE}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} (y_i - m\,x_i - b)^2\right) = -2\sum_{i=1}^{n} (y_i - m\,x_i - b)$$

Stationary values:

$$\frac{\partial\, SSE}{\partial b} = -2\sum_{i=1}^{n} (y_i - m\,x_i - b) = 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n} (y_i - m\,x_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + \ldots + 1}_{\text{n times 1}} = n$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Stationary values:

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \quad b\,n \quad = 0$$

Now we divide by n:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - m\,\frac{1}{n}\sum_{i=1}^{n} x_i - b \quad = 0$$

And please remember some basic definitions you have learnt earlier …

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Stationary values:

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \quad b\,n \quad = 0$$

Now we divide by n:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - m \,\frac{1}{n}\sum_{i=1}^{n} x_i - b \quad = 0$$

Please remember the definition of mean values:

$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ ………… arithmetic mean of y

$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ ………… arithmetic mean of x

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Stationary values:

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \quad b\,n \quad = 0$$

Now we divide by n and replace the sums:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - m\,\frac{1}{n}\sum_{i=1}^{n} x_i - b = 0$$

$$\Rightarrow \qquad \bar{y} \quad - m \qquad \bar{x} \quad - b = 0$$

$$\Rightarrow \qquad \boxed{\bar{y} = m\,\bar{x} + b}$$

As there are two unknown parameters m and b, we need two equations to determine them.

Therefore the partial derivative with respect to m will be found now.

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

Partial derivative of the sum of squared errors with respect to the variable m (slope):

$$\frac{\partial\,\text{SSE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n} (y_i - m\,x_i - b)^2\right)$$

$$= \sum_{i=1}^{n} \left(\frac{\partial}{\partial m} (y_i - m\,x_i - b)^2\right)$$

$$= \sum_{i=1}^{n} \left(2\,(y_i - m\,x_i - b)\,(-x_i)\right)$$

$$= -2 \sum_{i=1}^{n} \left((y_i - m\,x_i - b)\,x_i\right)$$

$$= -2 \sum_{i=1}^{n} \left(x_i\,y_i - m\,x_i^{\,2} - b\,x_i\right)$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Sum of squared errors:

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

Partial derivative of the sum of squared errors with respect to the variable m:

$$\frac{\partial\,\mathrm{SSE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n} (y_i - m\,x_i - b)^2\right) = -2\sum_{i=1}^{n}\left(x_i\,y_i - m\,x_i^2 - b\,x_i\right)$$

Stationary values:

$$\frac{\partial\,\mathrm{SSE}}{\partial m} = -2\sum_{i=1}^{n}\left(x_i\,y_i - m\,x_i^2 - b\,x_i\right) = 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n}\left(x_i\,y_i - m\,x_i^2 - b\,x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i\,y_i - m\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0$$

Again a division by n is required to replace the last term by the arithmetic mean of x.

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

$$\sum_{i=1}^{n} x_i\, y_i - m \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0$$

Again a division by n is required to replace the last term by the arithmetic mean of x:

$$\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - m\, \frac{1}{n}\sum_{i=1}^{n} x_i^2 - b\, \frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - m\, \frac{1}{n}\sum_{i=1}^{n} x_i^2 - b\, \bar{x} = 0$$

And we find some nice expressions …

… of the sum of squares $x_i^2$, which is related to the variance or square of the standard deviation,

… and of the sum of the product values $x_i\, y_i$, which is related to the covariance,

to also replace them soon.

# Reformulation of the Variance

*(Umformung der Varianz)*

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

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(x_i - \bar{x})$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2\,\bar{x}\,x_i + \bar{x}^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\bar{x}\;\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2\;\frac{1}{n}\sum_{i=1}^{n} 1$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\bar{x}\,\bar{x} + \bar{x}^2\,\frac{1}{n}\,n$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\bar{x}^2 + \bar{x}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$\Rightarrow \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} x_i^2 = s_x^2 + \bar{x}^2}$$

# Reformulation of the Covariance

*(Umformung der Kovarianz)*

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

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - 2\,\bar{x}\bar{y} + \bar{x}\bar{y}$$

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

$$\Rightarrow \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} x_i y_i = s_{xy} + \bar{x}\bar{y}}$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Replacing the sum of squares $x_i^2$ (which is related to the variance) and of the sum of the product values $x_i\,y_i$ (which is related to the covariance) in our intermediate result:

$$\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i - m\,\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - b\,\bar{x} = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} = s_x^{\,2} + \bar{x}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i = s_{xy} + \bar{x}\,\bar{y}$$

$$\Rightarrow\quad s_{xy} + \bar{x}\,\bar{y} - m\,s_x^{\,2} - m\,\bar{x}^2 - b\,\bar{x} = 0$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

First intermediate result:

$$\bar{y} = m\,\bar{x} + b$$

Second intermediate result:

$$s_{xy} + \bar{x}\,\bar{y} - m\,s_x^2 - m\,\bar{x}^2 - b\,\bar{x} = 0$$

$$\Rightarrow \quad s_{xy} + \bar{x}\,(m\,\bar{x} + b) - m\,s_x^2 - m\,\bar{x}^2 - b\,\bar{x} = 0$$

$$s_{xy} + m\,\bar{x}^2 + b\,\bar{x} - m\,s_x^2 - m\,\bar{x}^2 - b\,\bar{x} = 0$$

$$s_{xy} \qquad - m\,s_x^2 \qquad = 0$$

$$\Rightarrow \qquad \boxed{m = \frac{s_{xy}}{s_x^2}}$$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Preliminary result:

The least-square line (regression line)

$$y = m\,x + b$$

has the slope $\quad m = \dfrac{s_{xy}}{s_x^{\,2}}$

and the y-intercept $\quad b = \bar{y} - m\,\bar{x} = \bar{y} - \dfrac{s_{xy}}{s_x^{\,2}}\,\bar{x}$

Therefore the parameters m and b are called "regression coefficients".

But we are not ready yet!
The second derivative test is still missing.

# Second Derivative Test

*(Überprüfung der zweiten Ableitungen)*

Sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - m\,x_i - b)^2$$

Partial derivatives of first order:

$$\frac{\partial\,SSE}{\partial b} = -2 \sum_{i=1}^{n} (y_i - m\,x_i - b)$$

$$\frac{\partial\,SSE}{\partial m} = -2 \sum_{i=1}^{n} \left(x_i\,y_i - m\,x_i^2 - b\,x_i\right)$$

Partial derivatives of second order:

$$\frac{\partial^2\,SSE}{\partial b^2} = 2 \sum_{i=1}^{n} 1 = 2\,n \qquad > \ 0$$

$$\frac{\partial^2\,SSE}{\partial m^2} = 2 \sum_{i=1}^{n} x_i^2 = 2\,n\,\overline{x_i^2} \qquad > \ 0$$

$$\frac{\partial^2\,SSE}{\partial b\,\partial m} = 2 \sum_{i=1}^{n} x_i = 2\,n\,\bar{x}$$

Minimum!

$$\Rightarrow \qquad 2\,n \cdot 2\,n\,\overline{x_i^2} > (2\,n\,\bar{x})^2$$

because $\overline{x_i^2} > \bar{x}^2$

# Finding the Least-Square Line

*(Bestimmung der Geraden kleinster Quadrate)*

Final result:

The least-square line (regression line)

$$y = m\,x + b$$

has the slope $$m = \frac{s_{xy}}{s_x^{\,2}}$$

and the y-intercept $$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_{xy}}{s_x^{\,2}}\,\bar{x}$$

Therefore the parameters m and b are called “regression coefficients”.

# Example Problem

*(Beispielaufgabe)*

The following data points are given:

| $x_i$ | $y_i$ |
| --- | --- |
| 2 | 4 |
| 3 | 5 |
| 4 | 5.5 |
| 5 | 5.5 |
| 6 | 6 |

a) Find the regression line (best-fit line) by minimizing the sum of errors in y-direction.

b) Find the regression line (best-fit line) by minimizing the sum of errors in x-direction.

Please compare both solutions and interpret the conceptual differences.

# Solution

*(Lösung)*

a) Table of data points:

| $x_i$ | $y_i$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 4 | 4 | 1.44 | 2.4 |
| 3 | 5 | 1 | 0.04 | 0.2 |
| 4 | 5.5 | 0 | 0.09 | 0 |
| 5 | 5.5 | 1 | 0.09 | 0.3 |
| 6 | 6 | 4 | 0.64 | 1.6 |
| 20 | 26 | 10 | b) | 4.5 |

$$\bar{x} = \frac{20}{5} = 4 \qquad s_x^2 = \frac{10}{5} = 2$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_{xy} = \frac{4.5}{5} = 0.9$$

Check:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{90}{5} - 4^2 = 2$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y} = \frac{108.5}{5} - 4 \cdot 5.2 = 0.9$$

# Solution

*(Lösung)*

$$\bar{x} = \frac{20}{5} = 4 \qquad s_x^{\;2} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_{xy} = \frac{4.5}{5} = 0.9$$

$\Rightarrow$ Slope of the regression line:

$$m = \frac{s_{xy}}{s_x^{\;2}} = \frac{0.9}{2} = 0.45$$

$\Rightarrow$ y-intercept of the regression line:

$$b = \bar{y} - m\,\bar{x} = 5.2 - 0.45 \cdot 4 = 3.40$$

The regression line (best-fit line) with respect to minimized squares of errors in y-direction is:

$$y = 0.45\,x + 3.40$$

# Solution

*(Lösung)*

b) We can use the structural symmetry between minimizing squares of errors in y-direction and minimizing squares of errors in x-direction.

Table of data points:

| $x_i$ | $y_i$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 4 | 4 | 1.44 | 2.4 |
| 3 | 5 | 1 | 0.04 | 0.2 |
| 4 | 5.5 | 0 | 0.09 | 0 |
| 5 | 5.5 | 1 | 0.09 | 0.3 |
| 6 | 6 | 4 | 0.64 | 1.6 |
| 20 | 26 | a) | 2.30 | 4.5 |

$$\bar{x} = \frac{20}{5} = 4 \qquad s_y^{\,2} = \frac{2.3}{5} = 0.46$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_{xy} = \frac{4.5}{5} = 0.9$$

Check: $$s_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \bar{y}^2 = \frac{137.5}{5} - 5.2^2 = 0.46$$

# Solution

*(Lösung)*

$$\bar{x} = \frac{20}{5} = 4 \qquad s_y^{\,2} = \frac{2.3}{5} = 0.46$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_{xy} = \frac{4.5}{5} = 0.9$$

$\Rightarrow$ Slope of the regression line:

$$m = \frac{s_{xy}}{s_y^{\,2}} = \frac{0.9}{0.46} = \frac{45}{23} \approx 1.9565$$

$\Rightarrow$ x-intercept of the regression line:

$$b = \bar{x} - m\,\bar{y} = 4 - \frac{45}{23} \cdot 5.2 = -\frac{142}{23} \approx -6.1739$$

The regression line (best-fit line) with respect to minimized squares of errors in x-direction is:

$$x = \frac{45}{23}\,y - \frac{142}{23}$$

or solved for y: $y = \frac{23}{45}\,x + \frac{142}{45} \approx$ 0.5111 x + 3.1556

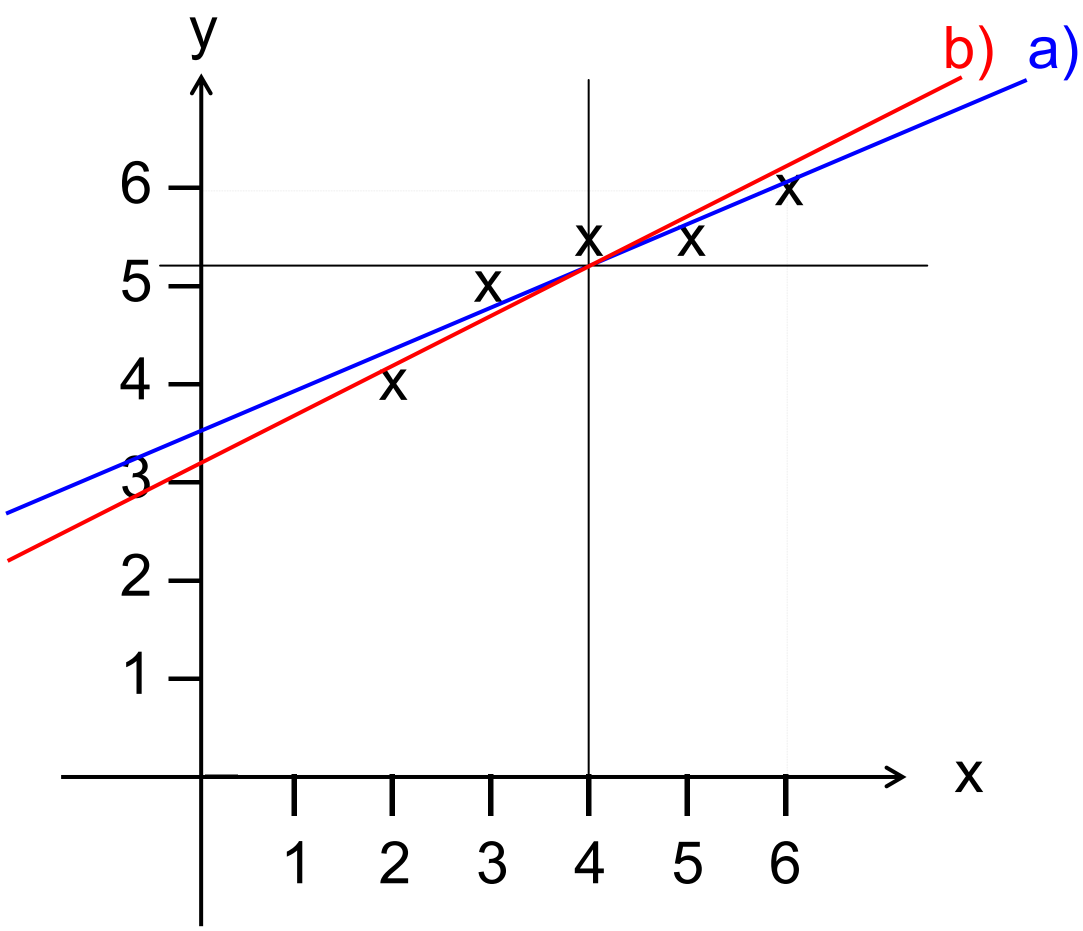
# Comparing both Solutions

*(Vergleich beider Lösung)*

a) $y = 0.45\,x + 3.40$ blue

b) $y = \frac{23}{45}\,x + \frac{142}{45} \approx 0.5111\,x + 3.1556$ red

Sketch of both regression lines:




Check of solutions:

$$y(4) = 0.45 \cdot 4 + 3.4 = \frac{23}{45} \cdot 4 + \frac{142}{45} = 5.2$$

# Interpretation of the Conceptual Differences

*(Interpretation der konzeptionellen Unterschiede)*

To interpret the conceptual differences between the solutions of part a) and part b), please have a look at the book

> Jens K. Perret: Arbeitsbuch zur Statistik für Wirtschafts- und Sozialwissenschaftler. Theorie, Aufgaben und Lösungen. Springer Gabler / Springer Fachmedien, Wiesbaden 2019

at section 7.1.2. ‘Lineare Regression’ on page 417:

> Bei der linearen Regression wird ein kausaler Effekt unterstellt. Es wird implizit angenommen, dass der Regressor (unabhängige Variable x) einen Einfluss auf den Regressanden (abhängige Variable y) ausübt, dass der Regressand sich somit aus den Regressoren ergibt.

Preliminary English translation by Prof. Perret:

> In linear regression, a causal effect is assumed. It is implicitly assumed that the regressor (independent variable x or feature) exerts an influence on the regressand (dependent variable y or label), that the regressand thus results from the regressors.

(Please see also section 6.1 ‘Mean Squared Error (MSE)’ on pp. 126/127 of the iubh course book “Advanced Mathematics”, DLMDSAM01, Version No. 001-2020-0824)

# Interpretation of the Conceptual Differences

*(Interpretation der konzeptionellen Unterschiede)*

Thus in part a) the variable x is the regressor (or feature) and variable y is the regressand (or label). It is assumed that variable x exerts an influence on variable y.

In part b) the variable y is the regressor and variable x is the regressand, and it is assumed that variable y exerts an influence on variable x.

But all these assumption about causal effects very often do not make much sense. We very often simply do not know how the causality of a situation should be described.

And we very often make mistakes and have to take observational errors into account not only when measuring y.

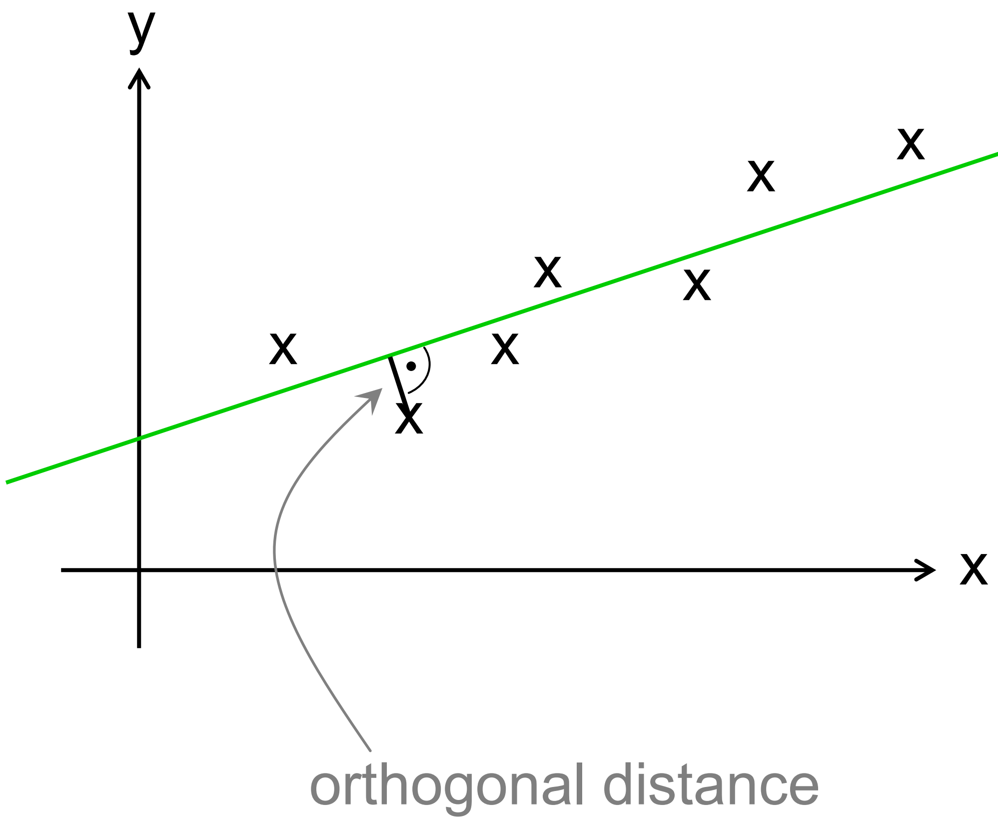
We very probably make mistakes and have to take observational errors into account when measuring the variable x and we make mis-takes and have to take observational errors into account when measuring the variable y.

# Orthogonal Regression

*(Orthogonale Regression)*

As in many situations observational errors and mistake in measuring of both variables occur, it makes sense to replace the traditional least squares method by orthogonal regression.

We will then minimize the square of the orthogonal distances to the regression line:



# Orthogonal Regression

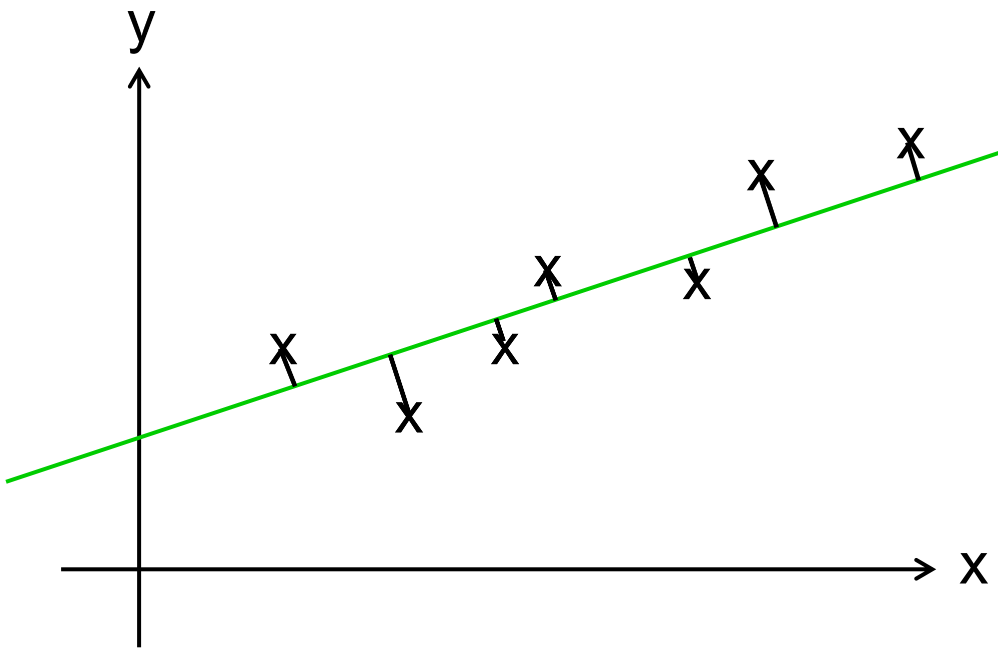
*(Orthogonale Regression)*

The new least-square line (new regression line) will be written again as:

$$y = m\,x + b$$

slope of the new regression line ($m$)

y-intercept of the new regression line ($b$)




Actual values of the data points:

$(x_1;\, y_1)$ $(x_2;\, y_2)$ $(x_3;\, y_3)$ $(x_4;\, y_4)$ …

Estimated values of the data points on the new regression line:

$(x_1^*;\, y_1^*)$ $(x_2^*;\, y_2^*)$ $(x_3^*;\, y_3^*)$ $(x_4^*;\, y_4^*)$ …

# Orthogonal Regression

*(Orthogonale Regression)*

Now both values change: the values of the x-coordinates and the values of the y-coordinates of the estimated values of the data points (foot points) on the new regression line:

measured data point: $(x_i; y_i)$

$\downarrow$

linear regression: $(x_i; \hat{y}_i)$

$\downarrow \quad \downarrow$

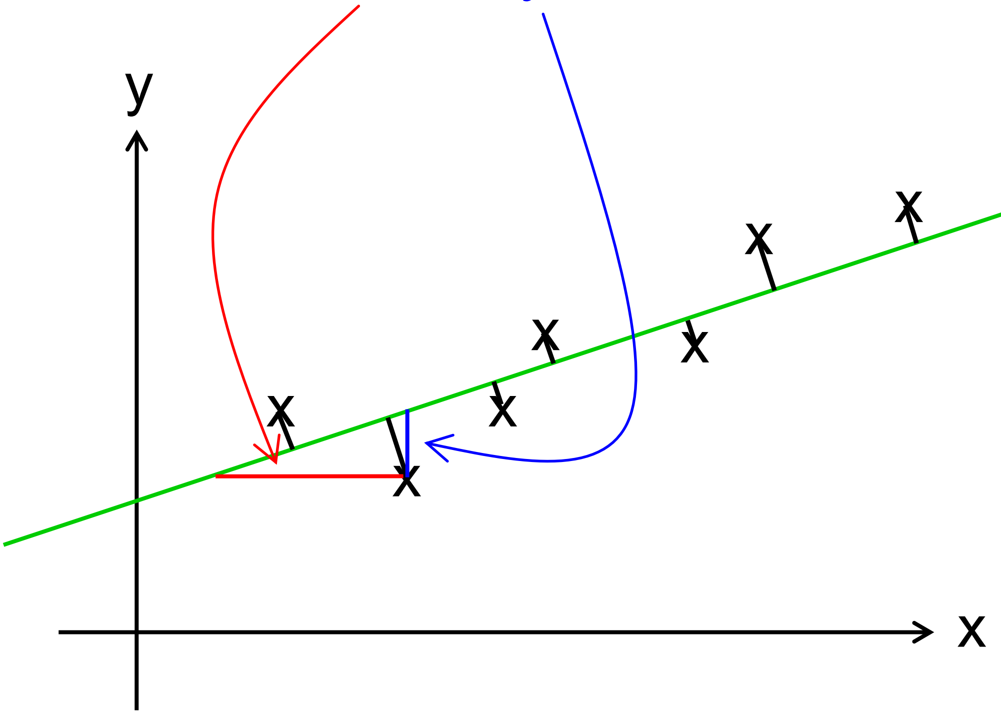
orthogonal regression: $(x_i^*; y_i^*)$

But already Pythagoras and Euclid have told us, how to replace the new values values $x_i^*$ and $y_i^*$ by the old values of $x_i$ and $\hat{y}_i$ when describing the orthogonal distance.

# Orthogonal Regression

*(Orthogonale Regression)*

Now we have to combine the two residuals into the directions of x and y …




Residuals (errors) of the data points:

$$e_{1x} = x_1 - x_1^*\qquad e_{1y} = y_1 - y_1^*$$

$$e_{2x} = x_2 - x_2^* \qquad e_{2y} = y_2 - y_2^*$$

$$e_{3x} = x_3 - x_3^* \qquad e_{3y} = y_3 - y_3^*$$

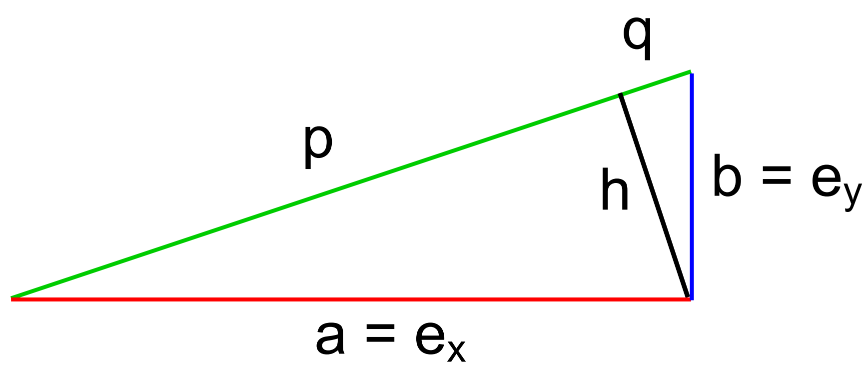
$$e_{4x} = x_4 - x_4^* \qquad e_{4y} = y_4 - y_4^* \qquad \text{etc …}$$

These residuals are the legs of right-angled triangles.

# Orthogonal Regression

*(Orthogonale Regression)*

Now we have to combine the two residuals into the directions of x and y to find the height of the right-angled triangle according to the theorems of Pythagoras and Euclid:




Euclid: $h^2 = p\,q$ $\qquad a^2 = e_x^{\,2} = p\,c = p\,(p + q)$

$$b^2 = e_y^{\,2} = q\,c = q\,(p + q)$$

$$\Rightarrow \qquad a^2\,b^2 = e_x^{\,2}\,e_y^{\,2} = p\,q\,c^2$$

$$\Rightarrow \qquad h^2 = \frac{e_x^{\,2}\,e_y^{\,2}}{c^2}$$

Pythagoras: $a^2 + b^2 = e_x^{\,2} + e_y^{\,2} = c^2 = (p + q)^2$

$$\Rightarrow \qquad h^2 = \frac{e_x^{\,2}\,e_y^{\,2}}{e_x^{\,2} + e_y^{\,2}}$$

# Orthogonal Regression

*(Orthogonale Regression)*

$$\Rightarrow \qquad h^2 = \frac{{e_x}^2\,{e_y}^2}{{e_x}^2 + {e_y}^2}$$

As the slope of the new regression line is given by

$$m = \frac{e_y}{e_x} \qquad \Rightarrow \qquad e_y = m\, e_x$$

the square of the height can be rewritten as:

$$h^2 = \frac{m^2}{1+m^2}\,{e_x}^2 = \frac{{e_y}^2}{1+m^2}$$

Thus we get the squares of these heights as squared new residuals (squared orthogonal errors) of the data points:

$${h_1}^2 = \frac{{e_{1y}}^2}{1+m^2} = \frac{(y_1 - \hat{y}_1)^2}{1+m^2} = \frac{(y_1 - m\,x_1 - b)^2}{1+m^2}$$

$${h_2}^2 = \frac{{e_{2y}}^2}{1+m^2} = \frac{(y_2 - \hat{y}_2)^2}{1+m^2} = \frac{(y_2 - m\,x_2 - b)^2}{1+m^2}$$

$${h_3}^2 = \frac{{e_{3y}}^2}{1+m^2} = \frac{(y_3 - \hat{y}_3)^2}{1+m^2} = \frac{(y_3 - m\,x_3 - b)^2}{1+m^2}$$

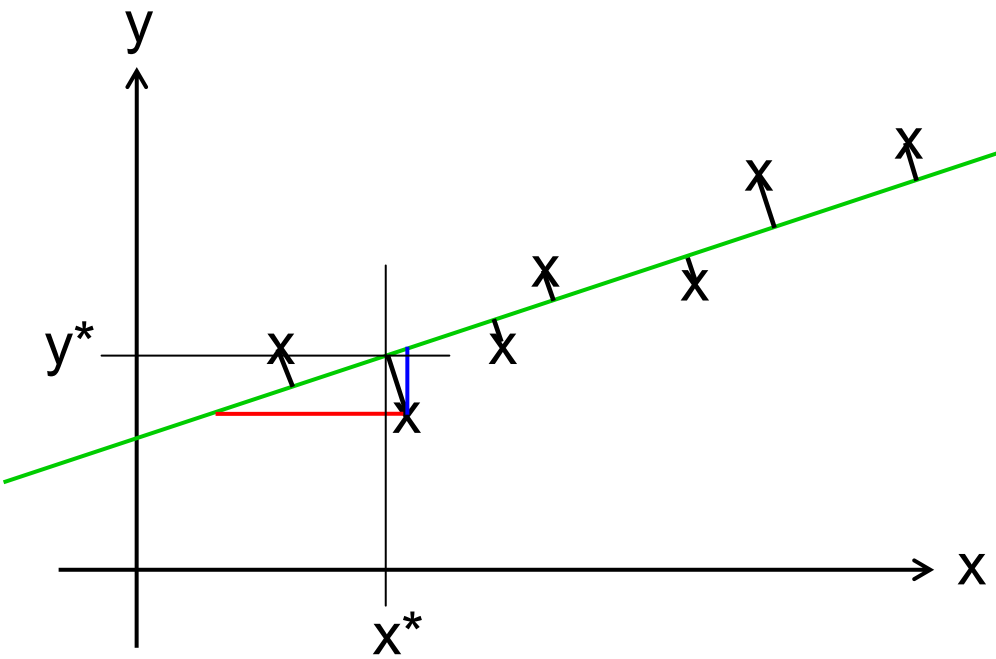
# What’s behind Pythagoras?

*(Was steckt hinter Pythagoras?)*

This result based on the rules of Pythagoras and Euclid can also be found by using optimization: We can minimize the sum of the squares of the orthogonal errors …

$$h^2 = e_x^2 + e_y^2 = (x - x^*)^2 + (y - y^*)^2$$

$$= (x - x^*)^2 + (y - m\,x^* - b)^2$$




… and will get an identical intermediate result at the end.

# What's behind Pythagoras?

*(Was steckt hinter Pythagoras?)*

The first derivative of the distance from a data point to the foot point at the new regression line will be:

$$\frac{d\,h^2}{dx^*} = \frac{d}{dx^*}\left[(x - x^*)^2 + (y - m\,x^* - b)^2\right]$$

$$= -2\,(x - x^*) - 2\,m\,(y - m\,x^* - b)$$

$$= -2\,(x - x^* + m\,y - m^2\,x^* - m\,b)$$

Thus we get the following stationary value:

$$-2\,(x - x^* + m\,y - m^2\,x^* - m\,b) = 0$$

$$x - x^* + m\,y - m^2\,x^* - m\,b = 0$$

$$\Rightarrow \qquad x^* = \frac{x + m\,y - m\,b}{1 + m^2}$$

# What's behind Pythagoras?

*(Was steckt hinter Pythagoras?)*

Second derivative test:

$$\frac{d^2 h^2}{dx^{*2}} = \frac{d^2}{dx^{*2}} \left[(x - x^*)^2 + (y - m x^* - b)^2\right]$$

$$= \frac{d}{dx^*} \left[- 2 (x - x^* + m y - m^2 x^* - m b)\right]$$

$$= - 2 (- 1 - m^2)$$

$$= 2 + 2 m^2 > 0$$

$\Rightarrow$ This is indeed a minimum.

We have found the minimized orthogonal distance. And the x-coordinate of the estimated value will be:

$$x^* = \frac{x + m y - m b}{1 + m^2}$$

# What's behind Pythagoras?

*(Was steckt hinter dem Pythagoras?)*

Substituting x* into the square of the height now results in:

$$h^2 = (x - x^*)^2 + (y - y^*)^2$$

$$= (x - x^*)^2 + (y - m\,x^* - b)^2$$

$$= \left(x - \frac{x + m\,y - m\,b}{1+m^2}\right)^2 + \left(y - m\,\frac{x + m\,y - m\,b}{1+m^2} - b\right)^2$$

$$= \left(\frac{m^2\,x - m\,y + m\,b}{1+m^2}\right)^2 + \left(y + \frac{-m\,x - m^2\,y + m^2\,b}{1+m^2} - b\right)^2$$

$$= \left(\frac{-m}{1+m^2}(y - m\,x - b)\right)^2 + \left(\frac{y - m\,x - b}{1+m^2}\right)^2$$

$$= \left(1+m^2\right)\left(\frac{y - m\,x - b}{1+m^2}\right)^2$$

$$= \frac{(y - m\,x - b)^2}{1+m^2} = \frac{{e_y}^2}{1+m^2}$$

This result is identical to the result, we have reached earlier by applying the theorems of Euclid and Pythagoras (see again following slide).

# Orthogonal Regression

*(Orthogonale Regression)*

$$\Rightarrow \qquad h^2 = \frac{e_x^{\,2}\, e_y^{\,2}}{e_x^{\,2} + e_y^{\,2}}$$

As the slope of the new regression line is given by

$$m = \frac{e_y}{e_x} \qquad \Rightarrow \qquad e_y = m\, e_x$$

the square of the height can be rewritten as:

$$h^2 = \frac{m^2}{1+m^2}\, e_x^{\,2} = \frac{e_y^{\,2}}{1+m^2}$$

Thus we get the squares of these heights as squared new residuals (squared orthogonal errors) of the data points:

$$h_1^{\,2} = \frac{e_{1y}^{\,2}}{1+m^2} = \frac{(y_1 - \hat{y}_1)^2}{1+m^2} = \frac{(y_1 - m\,x_1 - b)^2}{1+m^2}$$

$$h_2^{\,2} = \frac{e_{2y}^{\,2}}{1+m^2} = \frac{(y_2 - \hat{y}_2)^2}{1+m^2} = \frac{(y_2 - m\,x_2 - b)^2}{1+m^2}$$

$$h_3^{\,2} = \frac{e_{3y}^{\,2}}{1+m^2} = \frac{(y_3 - \hat{y}_3)^2}{1+m^2} = \frac{(y_3 - m\,x_3 - b)^2}{1+m^2}$$

etc …

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Sum of squared orthogonal errors (SSOE),
Sum of squared heights:

$$\text{SSOE} = \sum_{i=1}^{n} h_i^2 = h_1^2 + h_2^2 + h_3^2 + \ldots + h_n^2$$

$$= \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{1+m^2}$$

$$= \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}$$

The new least-square line is the line at which the sum of squared orthogonal errors has a minimum.

The values of the data points $(x_i; y_i)$ are given because they had been measured. Thus the sum of squared orthogonal errors (SSOE) depends on the two variables m and b.

⇒ We have to find the minimum of the sum of squared orthogonal errors (SSOE) with respect to these two variables m and b.

⇒ The partial derivatives of the sum of squared orthogonal errors with respect to m and b are required.

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Sum of squared orthogonal errors:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}$$

Partial derivative of the sum of squared ortho-gonal errors with respect to the variable b (y-intercept):

$$\frac{\partial\,\text{SSOE}}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}\right)$$

$$= \sum_{i=1}^{n} \left(\frac{\partial}{\partial b}\,\frac{(y_i - m\,x_i - b)^2}{1+m^2}\right)$$

$$= \sum_{i=1}^{n} \frac{2\,(y_i - m\,x_i - b)\,(-1)}{1+m^2}$$

$$= -2\sum_{i=1}^{n} \frac{y_i - m\,x_i - b}{1+m^2}$$

$$= -\frac{2}{1+m^2}\sum_{i=1}^{n} (y_i - m\,x_i - b)$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Sum of squared orthogonal errors:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}$$

Partial derivative of the sum of squared errors with respect to the variable b:

$$\frac{\partial\,\text{SSOE}}{\partial b} = -\frac{2}{1+m^2}\sum_{i=1}^{n}(y_i - m\,x_i - b)$$

Stationary values:

$$\frac{\partial\,\text{SSOE}}{\partial b} = -\frac{2}{1+m^2}\sum_{i=1}^{n}(y_i - m\,x_i - b) = 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n}(y_i - m\,x_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + \ldots + 1}_{\text{n times 1}} = n$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Stationary values:

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \quad b\, n \quad = 0$$

Now we divide by n:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - m \,\frac{1}{n}\sum_{i=1}^{n} x_i - b \quad = 0$$

And substituting …

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$ ………… arithmetic mean of y

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$ ………… arithmetic mean of x

… will result in:

$$\bar{y} - m\,\bar{x} - b = 0$$

$\Rightarrow$

$$\boxed{\bar{y} = m\,\bar{x} + b}$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Sum of squared orthogonal errors:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}$$

Now the partial derivative of the sum of squared orthogonal errors with respect to the variable m (slope) will be found:

$$\frac{\partial\,\text{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1+m^2}\right)$$

$$= \sum_{i=1}^{n} \left(\frac{\partial}{\partial m}\,\frac{(y_i - m\,x_i - b)^2}{1+m^2}\right)$$

$$= \sum_{i=1}^{n} \frac{2\,(1+m^2)\,(y_i - m\,x_i - b)\,(-x_i) - (y_i - m\,x_i - b)^2\,(2\,m)}{(1+m^2)^2}$$

$$= \frac{-2}{(1+m^2)^2} \sum_{i=1}^{n} \left((x_i + m^2\,x_i)\,(y_i - m\,x_i - b) + m\,(y_i - m\,x_i - b)^2\right)$$

$$= \frac{-2}{(1+m^2)^2} \sum_{i=1}^{n} \left((x_i + m^2\,x_i + m\,y_i - m^2\,x_i - m\,b)\,(y_i - m\,x_i - b)\right)$$

$$= \frac{-2}{(1+m^2)^2} \sum_{i=1}^{n} \left((x_i + m\,y_i - m\,b)\,(y_i - m\,x_i - b)\right)$$

$$= \frac{-2}{(1+m^2)^2} \sum_{i=1}^{n} (x_i y_i - m x_i^2 - b x_i + m y_i^2 - m^2 x_i y_i - mb y_i$$

$$- mb y_i + m^2 b x_i + mb^2)$$

$$= \frac{-2}{(1+m^2)^2} \sum_{i=1}^{n} (x_i y_i - m x_i^2 - b x_i + m y_i^2 - m^2 x_i y_i$$

$$- 2mb y_i + m^2 b x_i + mb^2)$$

$$= \frac{2}{(1+m^2)^2} \sum_{i=1}^{n} ((1-m^2) b x_i + 2mb y_i - (1-m^2) x_i y_i$$

$$+ m x_i^2 - m y_i^2 - mb^2)$$

$$= \frac{2b(1-m^2)}{(1+m^2)^2} \sum_{i=1}^{n} x_i + \frac{4mb}{(1+m^2)^2} \sum_{i=1}^{n} y_i - \frac{2(1-m^2)}{(1+m^2)^2} \sum_{i=1}^{n} x_i y_i$$

$$+ \frac{2m}{(1+m^2)^2} \sum_{i=1}^{n} x_i^2 - \frac{2m}{(1+m^2)^2} \sum_{i=1}^{n} y_i^2 - \frac{2mb^2}{(1+m^2)^2} \sum_{i=1}^{n} 1$$

$$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + \ldots + 1}_{\text{n times 1}} = n$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

$$\frac{\partial\,\mathrm{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n}\frac{(y_i - m\,x_i - b)^2}{1+m^2}\right)$$

$$= \frac{2\,b\,(1-m^2)}{(1+m^2)^2}\sum_{i=1}^{n}x_i + \frac{4\,m\,b}{(1+m^2)^2}\sum_{i=1}^{n}y_i - \frac{2\,(1-m^2)}{(1+m^2)^2}\sum_{i=1}^{n}x_i\,y_i$$

$$+ \frac{2\,m}{(1+m^2)^2}\sum_{i=1}^{n}x_i^{\,2} - \frac{2\,m}{(1+m^2)^2}\sum_{i=1}^{n}y_i^{\,2} - \frac{2\,m\,b^2}{(1+m^2)^2}\,n$$

Stationary values:

$$\frac{\partial\,\mathrm{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n}\frac{(y_i - m\,x_i - b)^2}{1+m^2}\right) = 0$$

$$0 = \frac{2\,b\,(1-m^2)}{(1+m^2)^2}\sum_{i=1}^{n}x_i + \frac{4\,m\,b}{(1+m^2)^2}\sum_{i=1}^{n}y_i - \frac{2\,(1-m^2)}{(1+m^2)^2}\sum_{i=1}^{n}x_i\,y_i$$

$$+ \frac{2\,m}{(1+m^2)^2}\sum_{i=1}^{n}x_i^{\,2} - \frac{2\,m}{(1+m^2)^2}\sum_{i=1}^{n}y_i^{\,2} - \frac{2\,m\,b^2}{(1+m^2)^2}\,n$$

Now we divide again by n …

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Now we divide again by n and multiply by the denominator $(1 + m^2)^2$ …

$$0 = 2\,b\,(1 - m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + 4\,m\,b\cdot\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$- 2\,(1 - m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i + 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$- 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\,m\,b^2$$

… and substitute:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = s_x^2 + \bar{x}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 = s_y^2 + \bar{y}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i = s_{xy} + \bar{x}\,\bar{y}$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

$$0 = 2\,b\,(1 - m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + 4\,m\,b\cdot\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$- 2\,(1 - m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i + 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$$

$$- 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - 2\,m\,b^2$$

$$0 = 2\,b\,(1 - m^2)\,\bar{x} + 4\,m\,b\,\bar{y} - 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^{\,2} + \bar{x}^2) - 2\,m\,(s_y^{\,2} + \bar{y}^2) - 2\,m\,b^2$$

Together with the previous result

$$\bar{y} = m\,\bar{x} + b$$

it should be possible to solve these two equations for m and b.

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

$$0 = 2\,b\,(1 - m^2)\,\bar{x} + 4\,m\,b\,\bar{y} - 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^{\,2} + \bar{x}^2) - 2\,m\,(s_y^{\,2} + \bar{y}^2) - 2\,m\,b^2$$

$$0 = 2\,b\,(1 - m^2)\,\bar{x} + 4\,m^2\,b\,\bar{x} + 4\,m\,b^2$$
$$- 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^{\,2} + \bar{x}^2) - 2\,m\,(s_y^{\,2} + \bar{y}^2) - 2\,m\,b^2$$

$$0 = 2\,b\,(1 + m^2)\,\bar{x} - 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^{\,2} + \bar{x}^2) - 2\,m\,(s_y^{\,2} + \bar{y}^2) + 2\,m\,b^2$$

$$\bar{y} - m\,\bar{x} = b \qquad \bar{y}^2 - 2\,m\,\bar{x}\,\bar{y} + m^2\,\bar{x}^2 = b^2$$

$$0 = 2\,(1 + m^2)\,\bar{x}\,\bar{y} - 2\,m\,(1 + m^2)\,\bar{x}^2$$
$$- 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^{\,2} + \bar{x}^2) - 2\,m\,(s_y^{\,2} + \bar{y}^2)$$
$$+ 2\,m\,\bar{y}^2 - 4\,m^2\,\bar{x}\,\bar{y} + 2\,m^3\,\bar{x}^2$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

$$0 = 2\,(1 + m^2)\,\bar{x}\,\bar{y} - 2\,m\,(1 + m^2)\,\bar{x}^2$$
$$- 2\,(1 - m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^2 + \bar{x}^2) - 2\,m\,(s_y^2 + \bar{y}^2)$$
$$+ 2\,m\,\bar{y}^2 - 4\,m^2\,\bar{x}\,\bar{y} + 2\,m^3\,\bar{x}^2$$

Some terms can be cancelled …

$$0 = 4\,m^2\,\bar{x}\,\bar{y} - 2\,(1 - m^2)\,s_{xy} + 2\,m\,s_x^2$$
$$- 2\,m\,(s_y^2 + \bar{y}^2) + 2\,m\,\bar{y}^2 - 4\,m^2\,\bar{x}\,\bar{y}$$

… and cancelled …

$$0 = -\,2\,(1 - m^2)\,s_{xy} + 2\,m\,s_x^2 - 2\,m\,s_y^2$$

$$\Rightarrow\ 0 = (1 - m^2)\,s_{xy} - m\,s_x^2 + m\,s_y^2$$

This looks like a quadratic equation:

$$0 = s_{xy}\,(-\,m^2) + (s_y^2 - s_x^2)\,m + s_{xy}$$

$$0 = s_{xy}\,m^2 + (s_x^2 - s_y^2)\,m - s_{xy}$$

$$\Rightarrow\ 0 = m^2 + \frac{s_x^2 - s_y^2}{s_{xy}}\,m - 1$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

After completing the square,

$$0 = m^2 + \frac{s_x^{\ 2} - s_y^{\ 2}}{s_{xy}}\ m - 1$$

$$0 = m^2 + \frac{s_x^{\ 2} - s_y^{\ 2}}{s_{xy}}\ m + \frac{\left(s_x^{\ 2} - s_y^{\ 2}\right)^2}{4\, s_{xy}^{\ 2}} - \frac{\left(s_x^{\ 2} - s_y^{\ 2}\right)^2}{4\, s_{xy}^{\ 2}} - 1$$

$$0 = \left(m + \frac{s_x^{\ 2} - s_y^{\ 2}}{2\, s_{xy}}\right)^2 - \frac{\left(s_x^{\ 2} - s_y^{\ 2}\right)^2}{4\, s_{xy}^{\ 2}} - 1$$

we get:

$$\left(m + \frac{s_x^{\ 2} - s_y^{\ 2}}{2\, s_{xy}}\right)^2 = \frac{\left(s_x^{\ 2} - s_y^{\ 2}\right)^2}{4\, s_{xy}^{\ 2}} + 1$$

$$= \frac{\left(s_x^{\ 2} - s_y^{\ 2}\right)^2 + 4\, s_{xy}^{\ 2}}{4\, s_{xy}^{\ 2}}$$

$$m + \frac{s_x^{\ 2} - s_y^{\ 2}}{2\, s_{xy}} = \pm \frac{\sqrt{(s_x^{\ 2} - s_y^{\ 2})^2 + 4\, s_{xy}^{\ 2}}}{2\, s_{xy}}$$

$$m = \frac{s_y^{\ 2} - s_x^{\ 2} \pm \sqrt{(s_x^{\ 2} - s_y^{\ 2})^2 + 4\, s_{xy}^{\ 2}}}{2\, s_{xy}}$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Intermediate preliminary result:

The least-square line (regression line)

$$y = m\,x + b$$

has the slope
$$m = \frac{s_y^{\,2} - s_x^{\,2} \pm \sqrt{(s_x^{\,2} - s_y^{\,2})^2 + 4\,s_{xy}^{\,2}}}{2\,s_{xy}}$$

and the y-intercept

$$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_y^{\,2} - s_x^{\,2} \pm \sqrt{(s_x^{\,2} - s_y^{\,2})^2 + 4\,s_{xy}^{\,2}}}{2\,s_{xy}}\,\bar{x}$$

Again the parameters m and b can be called “regression coefficients”.

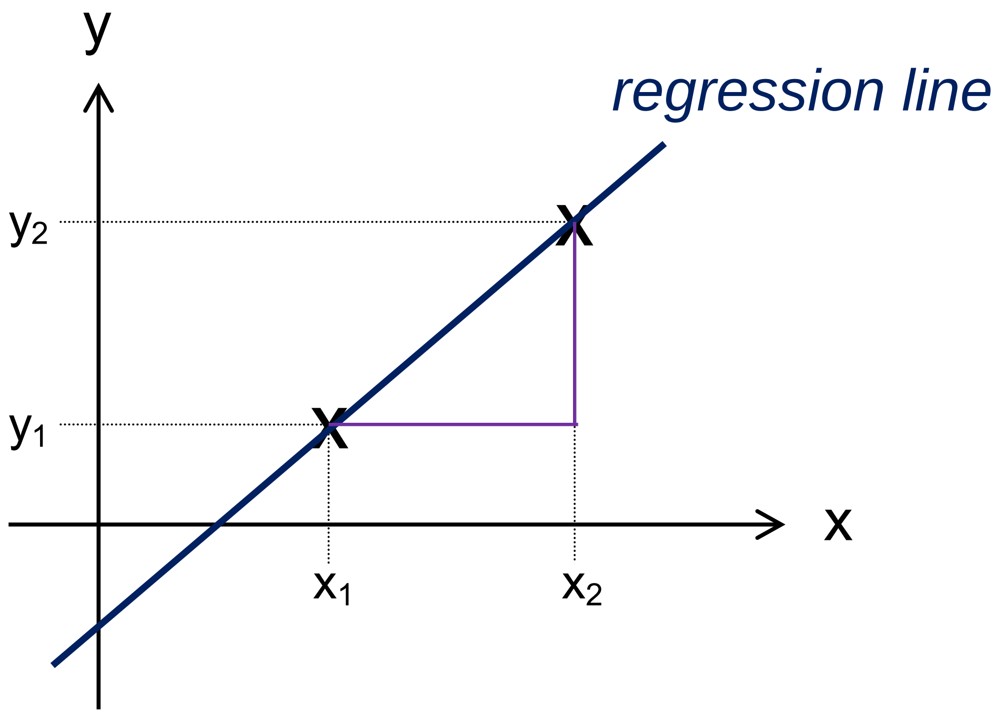
# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

We now have to decide whether the positive or the negative sign gives the correct answer. There is only one least-square line; thus only one of the signs can be the correct one.

Again it is of course possible to make a second derivative test. But we will follow a more direct strategy by simply checking the equation we have found in a very simple example:

If we measure only two data points $(x_1; y_1)$ and $(x_2; y_2)$, the errors have to disappear and the regression line will go through the two given points.



# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Thus we will have:

$$\left.\begin{array}{l}\Delta x = x_2 - x_1 \\ \Delta y = y_2 - y_1\end{array}\right\} \quad m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

As we know the perfect value of the slope m, we can compare it with the values we get with our intermediate equation

$$m = \frac{s_y{}^2 - s_x{}^2 \pm \sqrt{(s_x{}^2 - s_y{}^2)^2 + 4\, s_{xy}{}^2}}{2\, s_{xy}}$$

We only have to find variances and covariance.

For example we can take the two data points (0; 0) and (1; 1) to get a slope of m = 1:

$\bar{x} = 0.5 \qquad \bar{y} = 0.5$

$s_x{}^2 = 0.25 \qquad s_y{}^2 = 0.25 \qquad s_{xy} = 0.25$

$$m = \frac{0.25 - 0.25 \pm \sqrt{(0.25 - 0.25)^2 + 4 \cdot 0.25^2}}{2 \cdot 0.25} = \pm 1$$

$\Rightarrow$ The sign must be positive to get the slope m = 1.

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Or a little bit more general:
We take the two data points (0; 0) and ($x_2$; $y_2$)
to get a slope of $m = \frac{y_2}{x_2}$:

$\bar{x} = 0.5\, x_2 \qquad \bar{y} = 0.5\, y_2$

$s_x^2 = 0.25\, x_2^2 \qquad s_y^2 = 0.25\, y_2^2 \qquad s_{xy} = 0.25\, x_2 y_2$

$$m = \frac{s_y^2 - s_x^2 \pm \sqrt{(s_x^2 - s_y^2)^2 + 4\, s_{xy}^2}}{2\, s_{xy}}$$

$$= \frac{0.25\, y_2^2 - 0.25\, x_2^2 \pm \sqrt{(0.25\, y_2^2 - 0.25\, x_2^2)^2 + 4 \cdot (0.25\, x_2 y_2)^2}}{2 \cdot 0.25\, x_2 y_2}$$

$$= \frac{0.25\, y_2^2 - 0.25\, x_2^2 \pm \sqrt{(0.25\, y_2^2 + 0.25\, x_2^2)^2}}{2 \cdot 0.25\, x_2 y_2}$$

$$= \frac{0.25\, y_2^2 - 0.25\, x_2^2 \pm (0.25\, y_2^2 + 0.25\, x_2^2)}{2 \cdot 0.25\, x_2 y_2}$$

$\Rightarrow$ The sign must be positive to get the slope

$$m = \frac{0.25\, y_2^2 - 0.25\, x_2^2 + 0.25\, y_2^2 + 0.25\, x_2^2}{2 \cdot 0.25\, x_2 y_2} = \frac{y_2}{x_2}$$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

The last slide can also help us to understand the meaning of the negative sign. If we chose the negative sign, we will get a slope of

$$m = \frac{s_y^2 - s_x^2 - \sqrt{(s_x^2 - s_y^2)^2 + 4\, s_{xy}^2}}{2\, s_{xy}}$$

$$= \frac{0.25\, y_2^2 - 0.25\, x_2^2 - 0.25\, y_2^2 - 0.25\, x_2^2}{2 \cdot 0.25\, x_2 y_2}$$

$$= -\frac{x_2}{y_2}$$

which is in general:

$$m = \frac{s_y^2 - s_x^2 - \sqrt{(s_x^2 - s_y^2)^2 + 4\, s_{xy}^2}}{2\, s_{xy}}$$

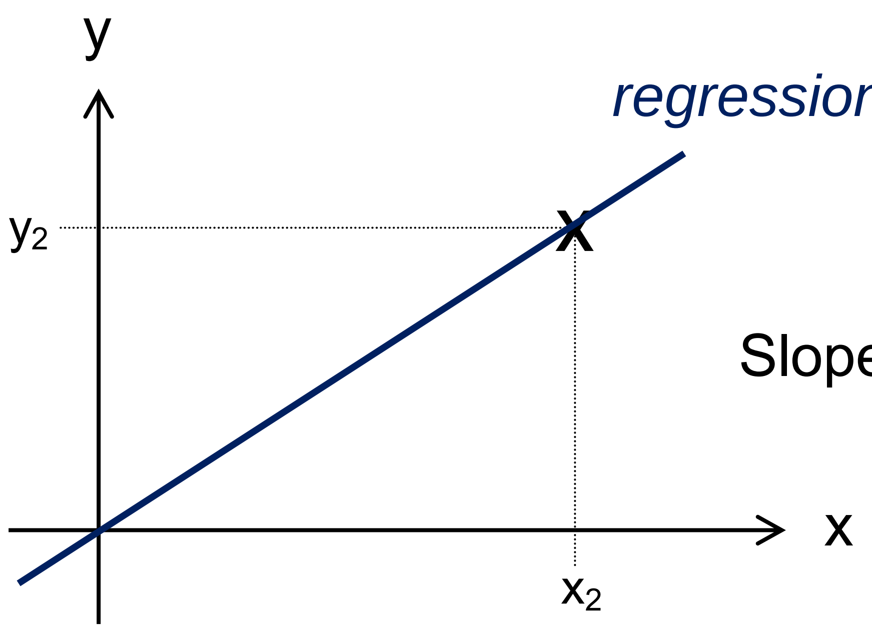
$$= -\frac{s_x^2 - s_y^2 + \sqrt{(s_x^2 - s_y^2)^2 + 4\, s_{xy}^2}}{2\, s_{xy}}$$

Thus two transformations happen at the same time:

- Interchange of x-axis and y-axis
- Change of handedness of the co-ordinate system

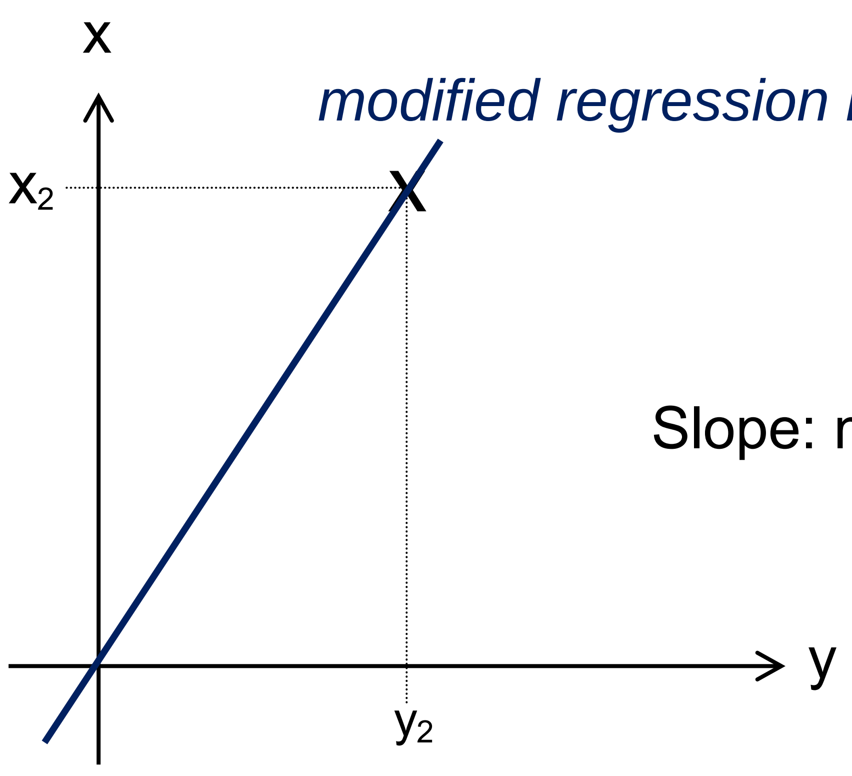
# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*




Slope: $m = \frac{y_2}{x_2}$

Interchange of x-axis and y-axis:

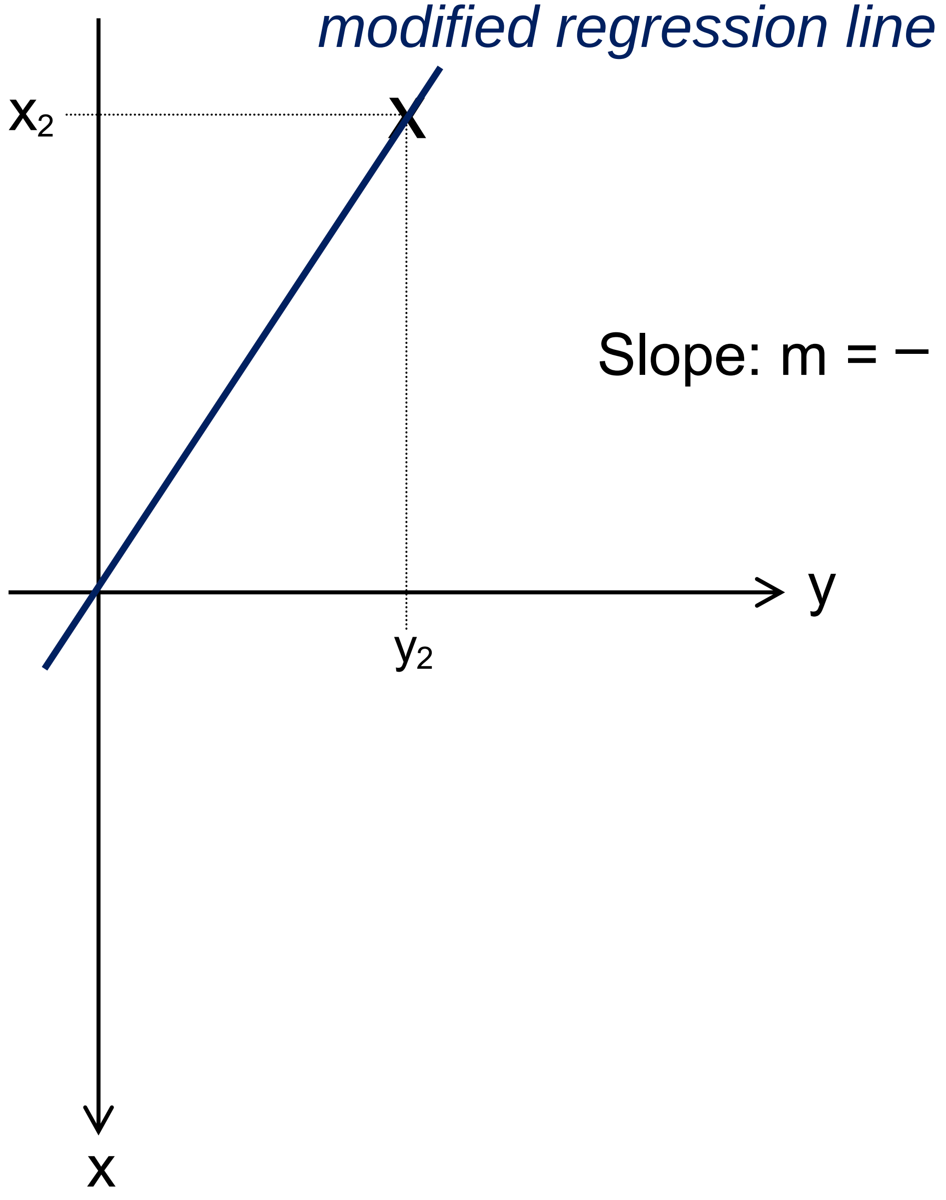



Slope: $m = \frac{x_2}{y_2}$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Change of handedness of the coordinate system:

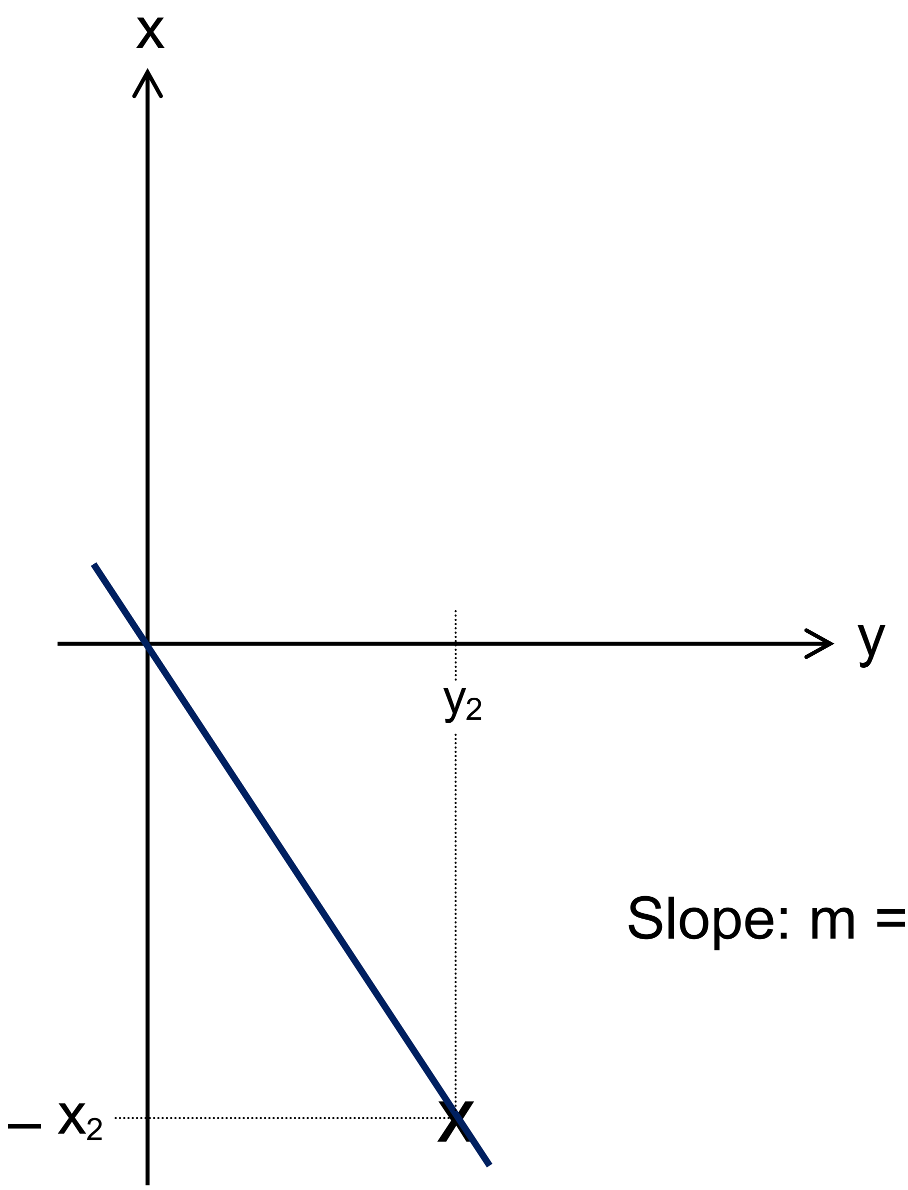



Slope: $m = -\frac{x_2}{y_2}$

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Or:




Slope: $m = \dfrac{-x_2}{y_2}$

*modified regression line*

# Finding the New Least-Square Line

*(Bestimmung der neuen Geraden kleinster Quadrate)*

Thus both solutions are meaningful. But when not caring about any trouble of interchanging axes or changing the handedness of the coordinate system, the following final result can be found in books about orthogonal regression:

The least-square line (regression line)

$$y = m\,x + b$$

has the slope

$$m = \frac{s_y{}^2 - s_x{}^2 + \sqrt{(s_x{}^2 - s_y{}^2)^2 + 4\,s_{xy}{}^2}}{2\,s_{xy}}$$

and the y-intercept

$$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_y{}^2 - s_x{}^2 + \sqrt{(s_x{}^2 - s_y{}^2)^2 + 4\,s_{xy}{}^2}}{2\,s_{xy}}\,\bar{x}$$

when orthogonal regression is applied.

# Modified example Problem

*(Modifizierte Beispielaufgabe)*

The following data points are given:

| $x_i$ | $y_i$ |
|---|---|
| 2 | 4 |
| 3 | 5 |
| 4 | 5.5 |
| 5 | 5.5 |
| 6 | 6 |

Find the regression line (best-fit line) by applying orthogonal regression.

# Solution

*(Lösung)*

Table of data points:

| $x_i$ | $y_i$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 4 | 4 | 1.44 | 2.4 |
| 3 | 5 | 1 | 0.04 | 0.2 |
| 4 | 5.5 | 0 | 0.09 | 0 |
| 5 | 5.5 | 1 | 0.09 | 0.3 |
| 6 | 6 | 4 | 0.64 | 1.6 |
| 20 | 26 | 10 | 2.30 | 4.5 |

$$\bar{x} = \frac{20}{5} = 4 \qquad s_x^2 = \frac{10}{5} = 2$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_y^2 = \frac{2.3}{5} = 0.46$$

$$s_{xy} = \frac{4.5}{5} = 0.9$$

# Solution

*(Lösung)*

Checks: $s_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2 = \frac{90}{5} - 4^2 = 2$

$$s_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \bar{y}^2 = \frac{137.5}{5} - 5.2^2 = 0.46$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y} = \frac{108.5}{5} - 4 \cdot 5.2 = 0.9$$

$\Rightarrow$ Slope of the orthogonal regression line:

$$m = \frac{s_y^{\,2} - s_x^{\,2} + \sqrt{(s_x^{\,2} - s_y^{\,2})^2 + 4\, s_{xy}^{\,2}}}{2\, s_{xy}}$$

$$= \frac{0.46 - 2 + \sqrt{(2 - 0.46)^2 + 4 \cdot 0.9^2}}{2 \cdot 0.9}$$

$$= \frac{-1.54 + \sqrt{1.54^2 + 3.24}}{1.8}$$

$$= \frac{-1.54 + \sqrt{5.6116}}{1.8} = \frac{0.8289}{1.8}$$

$$= 0.46049$$

$$\approx 0.46$$

# Solution

*(Lösung)*

$\Rightarrow$ y-intercept of the orthogonal regression line:

$$b = \bar{y} - m\,\bar{x}$$

$$= 5.2 - 0.46049 \cdot 4$$

$$= 3.35804$$

$$\approx 3.36$$

The orthogonal regression line (best-fit line) with respect to minimized squares of orthogonal errors will be:

$$y = 0.46049\,x + 3.35804$$

Comparison: The slope of this new orthogonal regression line lies between the slopes of the two linear regression lines found earlier.

$$y_{\text{linear-y}} = 0.45\,x + 3.40$$

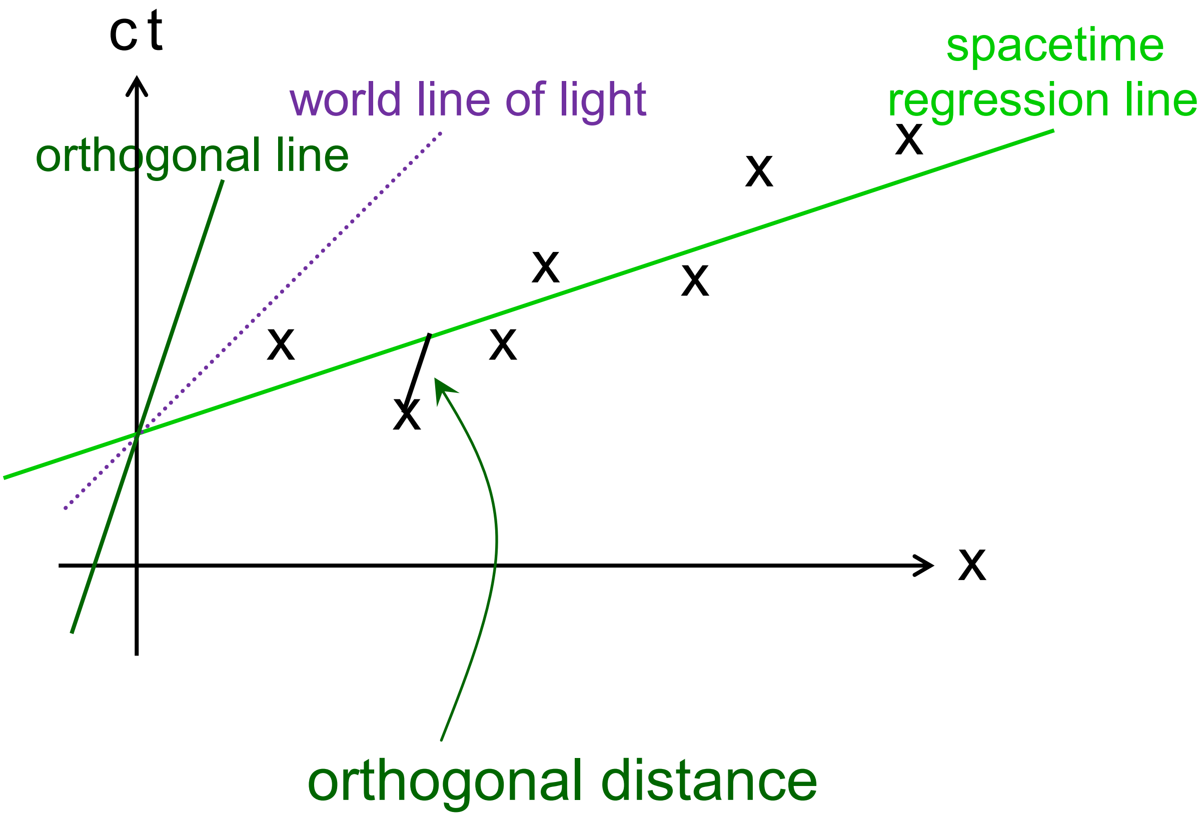
$$y_{\text{linear-x}} = \frac{23}{45}\,x + \frac{142}{45} \approx 0.5111\,x + 3.1556$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Now we will have a look at orthogonal regression in spacetime. Therefore the y-axis is replaced by a time axis into the direction of c t. As usual the velocity of light will be

$$c = 300\,000\ \frac{km}{s} = 3 \cdot 10^8\ \frac{m}{s}$$



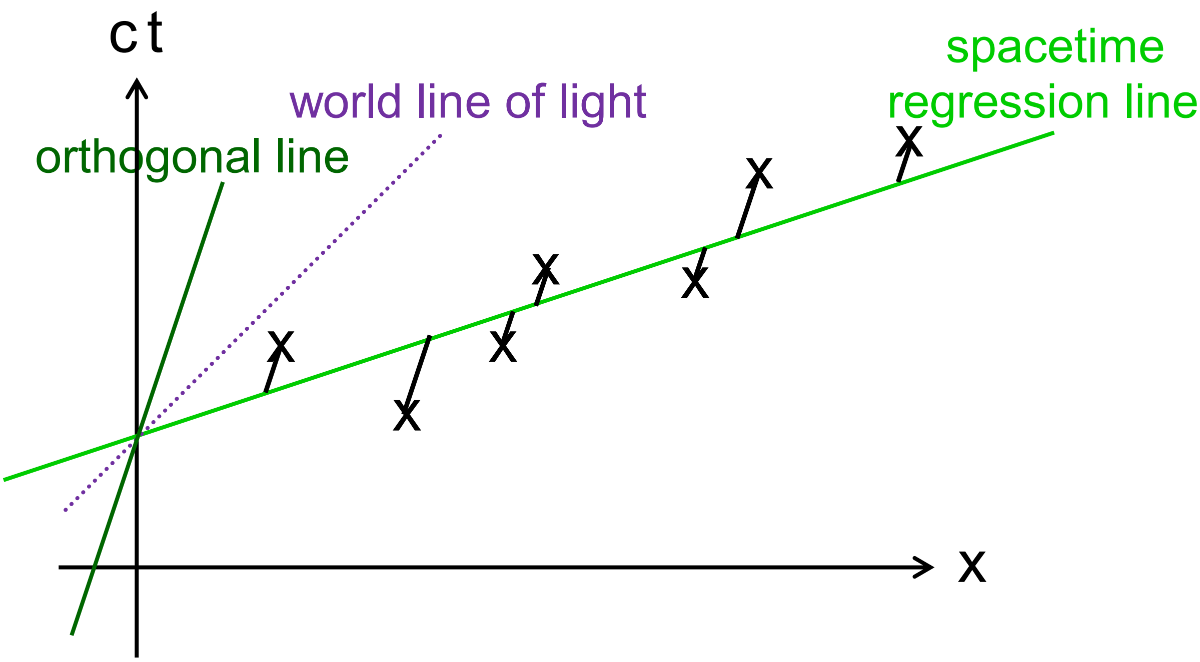
# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

The spacetime regression line will be written again as:

$$y = c\,t = m\,x + b$$

slope of the      y-intercept of the




Actual values of the data points:

$(x_1;\, y_1)$    $(x_2;\, y_2)$    $(x_3;\, y_3)$    $(x_4;\, y_4)$    . . .

or: $(x_1;\, c\,t_1)$    $(x_2;\, c\,t_2)$    $(x_3;\, c\,t_3)$    $(x_4;\, c\,t_4)$    . . .

Estimated values of the data points on the spacetime regression line:

$(x_1^+;\, y_1^+)$    $(x_2^+;\, y_2^+)$    $(x_3^+;\, y_3^+)$    $(x_4^+;\, y_4^+)$    . . .

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Again both values change: the values of the x-coordinates and the values of the y-coordinates of the estimated values of the data points (foot points) on the spacetime regression line:

measured data point: $(x_i; y_i)$

↓

linear regression: $(x_i; \hat{y}_i)$

↓ ↓

orthogonal regression: $(x_i^*; y_i^*)$

↓ ↓

orthogonal regression in spacetime: $(x_i^+; y_i^+) = (x_i^+; c\, t_i^+)$

But this time, the distance will not be minimized! Therefore we should not call the spacetime regression line a least-square line.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

We are now looking for an extreme value of the square of the distance:

$$h^2 = (e_t \gamma_t + e_x \gamma_x)^2$$

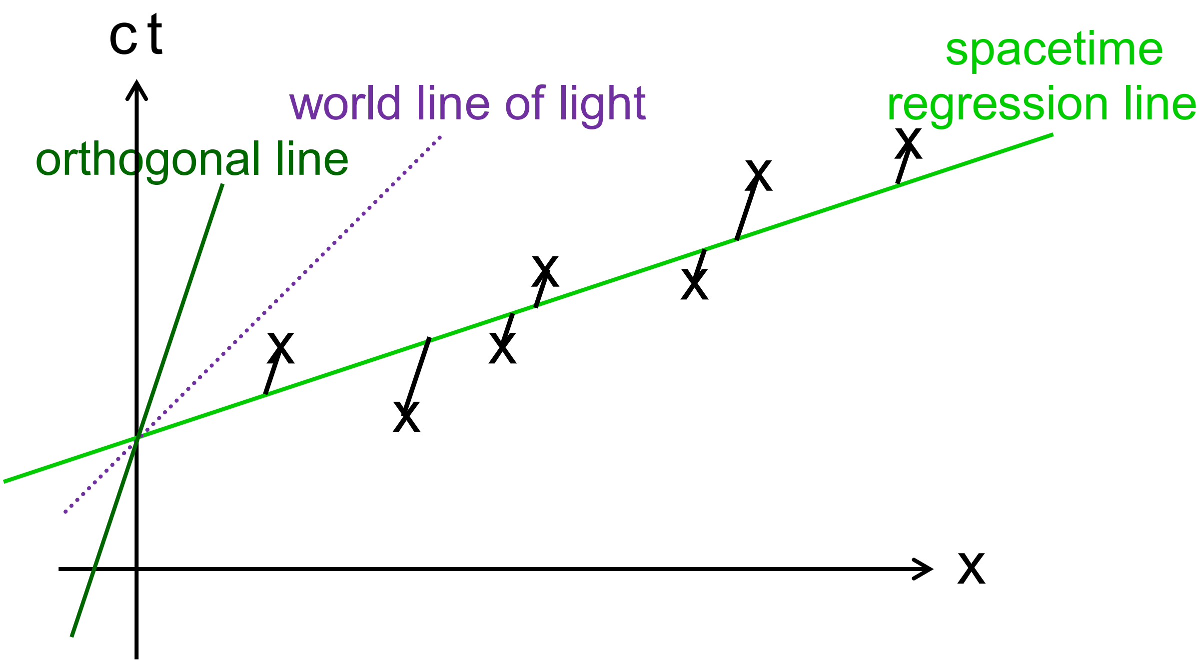
$$= e_y^2 \gamma_t^2 + e_x^2 \gamma_x^2 \qquad \text{with} \quad \gamma_t^2 = 1 \quad \gamma_x^2 = -1$$

(Spacetime Algebra)

$$= e_y^2 - e_x^2$$

$$= (y - y^+)^2 - (x - x^+)^2$$

$$= -(x - x^+)^2 + (y - m x^+ - b)^2$$




Spatial components square to negative values, because we have a hyperbolic metric now.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

The first derivative of the distance from a data point to the foot point at the spacetime regression line will be:

$$\frac{d\,h^2}{dx^+} = \frac{d}{dx^+}\left[-\left(x - x^+\right)^2 + \left(y - m\,x^+ - b\right)^2\right]$$

$$= 2\,(x - x^+) - 2\,m\,(y - m\,x^+ - b)$$

$$= 2\,(x - x^+ - m\,y + m^2\,x^+ + m\,b)$$

Thus we get the following stationary value:

$$2\,(x - x^+ - m\,y + m^2\,x^+ + m\,b) = 0$$

$$x - x^+ - m\,y + m^2\,x^+ + m\,b = 0$$

$$\Rightarrow \qquad x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

# <u>Orthogonal Regression in Spacetime</u>

*(Orthogonale Regression in der Raumzeit)*

Second derivative test:

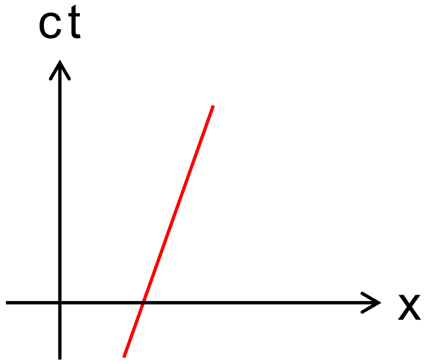
$$\frac{d^2 h^2}{dx^{+2}} = \frac{d^2}{dx^{+2}} [-(x - x^+)^2 + (y - m\,x^+ - b)^2]$$

$$= \frac{d}{dx^*} [2\,(x - x^+ - m\,y + m^2\,x^+ + m\,b)]$$

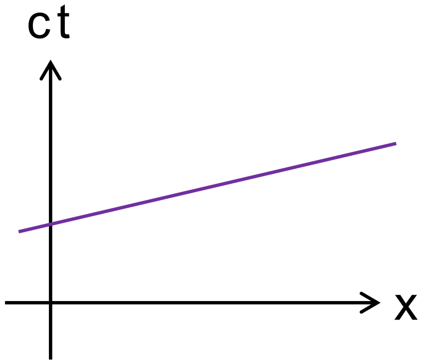
$$= 2\,(-1 + m^2)$$

$$= -2 + 2\,m^2$$

Objects, which exist in our world (the space-time world we actually live in) are moving with velocities smaller than the velocity of light. Thus their world lines in the Minkowski diagram sketched on the previous slide must have a slope which is greater than one:




world line of an object which exists in our world




world line of an hypothetical object moving with superluminal velocity

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Second derivative test:

$$\frac{d^2\, h^2}{dx^{+2}} = -2 + 2\, m^2$$

Thus there are two possible situations:

$m^2 > 1 \quad \Rightarrow \quad$ time-like regression line

$$\Rightarrow \quad \frac{d^2\, h^2}{dx^{+2}} = -2 + 2\, m^2 > 0$$

$\Rightarrow \quad$ minimum

The vector **h** of the orthogonal distance is a space-like vector and its square $h^2$ is negative. Thus the minimum is the greatest negative value with maximized length of the vector.

$m^2 < 1 \quad \Rightarrow \quad$ space-like regression line

$$\Rightarrow \quad \frac{d^2\, h^2}{dx^{+2}} = -2 + 2\, m^2 < 0$$

$\Rightarrow \quad$ maximum

The vector **h** of the orthogonal distance is a time-like vector and its square $h^2$ is positive. tive. Thus the length of **h** is maximized.

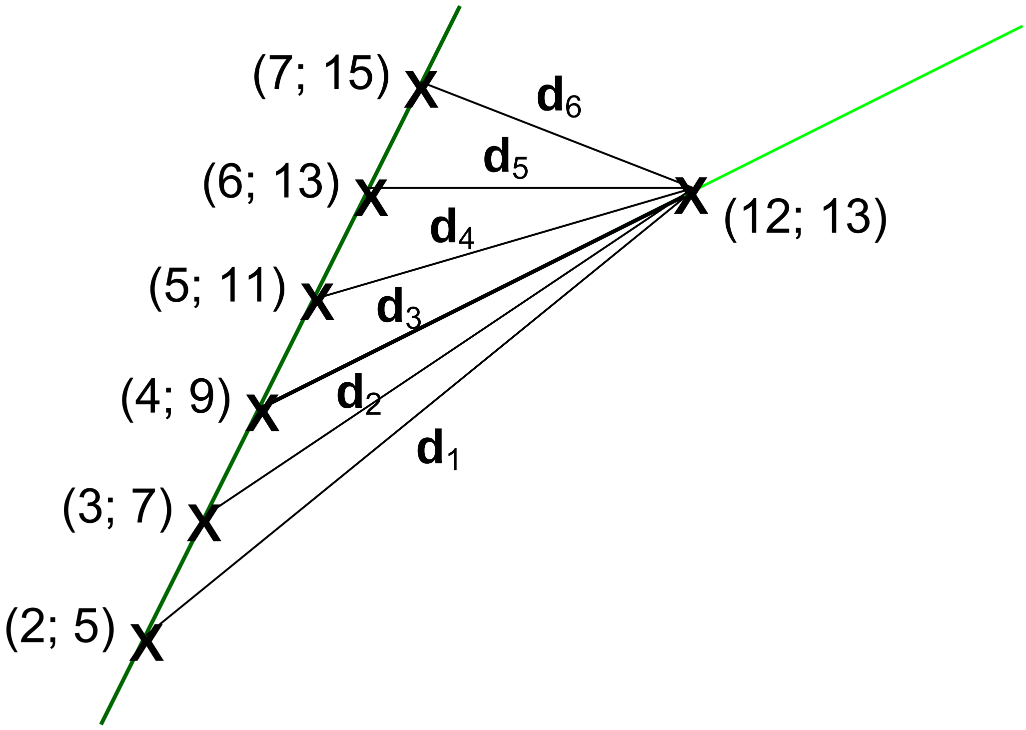
# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Some Examples:

(1) $y = c\,t = 2\,x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line




$$\mathbf{d}_6 = (12 - 7)\,\gamma_x + (13 - 15)\,\gamma_t = 5\,\gamma_x - 2\,\gamma_t$$

$$\mathbf{d}_5 = (12 - 6)\,\gamma_x + (13 - 13)\,\gamma_t = 6\,\gamma_x$$

$$\mathbf{d}_4 = (12 - 5)\,\gamma_x + (13 - 11)\,\gamma_t = 7\,\gamma_x + 2\,\gamma_t$$

$$\mathbf{d}_3 = (12 - 4)\,\gamma_x + (13 - 9)\,\gamma_t = 8\,\gamma_x + 4\,\gamma_t$$

$$\mathbf{d}_2 = (12 - 3)\,\gamma_x + (13 - 7)\,\gamma_t = 9\,\gamma_x + 6\,\gamma_t$$

$$\mathbf{d}_1 = (12 - 2)\,\gamma_x + (13 - 5)\,\gamma_t = 10\,\gamma_x + 8\,\gamma_t$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(1) $y = ct = 2x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line

$\mathbf{d}_6 = (12 - 7)\,\gamma_x + (13 - 15)\,\gamma_t = 5\,\gamma_x - 2\,\gamma_t$

$\mathbf{d}_5 = (12 - 6)\,\gamma_x + (13 - 13)\,\gamma_t = 6\,\gamma_x$

$\mathbf{d}_4 = (12 - 5)\,\gamma_x + (13 - 11)\,\gamma_t = 7\,\gamma_x + 2\,\gamma_t$

$\mathbf{d}_3 = (12 - 4)\,\gamma_x + (13 - 9)\,\gamma_t = 8\,\gamma_x + 4\,\gamma_t$

$\mathbf{d}_2 = (12 - 3)\,\gamma_x + (13 - 7)\,\gamma_t = 9\,\gamma_x + 6\,\gamma_t$

$\mathbf{d}_1 = (12 - 2)\,\gamma_x + (13 - 5)\,\gamma_t = 10\,\gamma_x + 8\,\gamma_t$

Squares of the distance vectors $\mathbf{d}_i$:

$\mathbf{d}_6^2 = (5\,\gamma_x - 2\,\gamma_t)^2 = -25 + 4 = -21$

$\mathbf{d}_5^2 = (6\,\gamma_x)^2 = -36$

$\mathbf{d}_4^2 = (7\,\gamma_x + 2\,\gamma_t)^2 = -49 + 4 = -45$

$\mathbf{d}_3^2 = (8\,\gamma_x + 4\,\gamma_t)^2 = -64 + 16 = -48$

$\mathbf{d}_2^2 = (9\,\gamma_x + 6\,\gamma_t)^2 = -81 + 36 = -45$

$\mathbf{d}_1^2 = (10\,\gamma_x + 8\,\gamma_t)^2 = -100 + 64 = -36$

Minimum of squares

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(1) $y = ct = 2x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line

The magnitude of the distance vectors $\mathbf{d}_i$ is identical to the length of the distance vectors:

$\mathbf{d}_6^2 = -21 \quad \Rightarrow \quad |\mathbf{d}_6^2| = 21 \quad \Rightarrow \quad |\mathbf{d}_6| = \sqrt{21}$

$\mathbf{d}_5^2 = -36 \quad \Rightarrow \quad |\mathbf{d}_5^2| = 36 \quad \Rightarrow \quad |\mathbf{d}_5| = \sqrt{36}$

$\mathbf{d}_4^2 = -45 \quad \Rightarrow \quad |\mathbf{d}_4^2| = 45 \quad \Rightarrow \quad |\mathbf{d}_4| = \sqrt{45}$

$\mathbf{d}_3^2 = -48 \quad \Rightarrow \quad |\mathbf{d}_3^2| = 48 \quad \Rightarrow \quad |\mathbf{d}_3| = \sqrt{48}$

$\mathbf{d}_2^2 = -45 \quad \Rightarrow \quad |\mathbf{d}_2^2| = 45 \quad \Rightarrow \quad |\mathbf{d}_2| = \sqrt{45}$

$\mathbf{d}_1^2 = -36 \quad \Rightarrow \quad |\mathbf{d}_1^2| = 36 \quad \Rightarrow \quad |\mathbf{d}_1| = \sqrt{36}$

The minimum of squares is identical to the maximum of the lengths.

We have indeed found the distance vector with a maximized magnitude of the square, when the squares are minimized!

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(1) $y = ct = 2x + 1 \quad \Rightarrow \quad m^2 > 1$

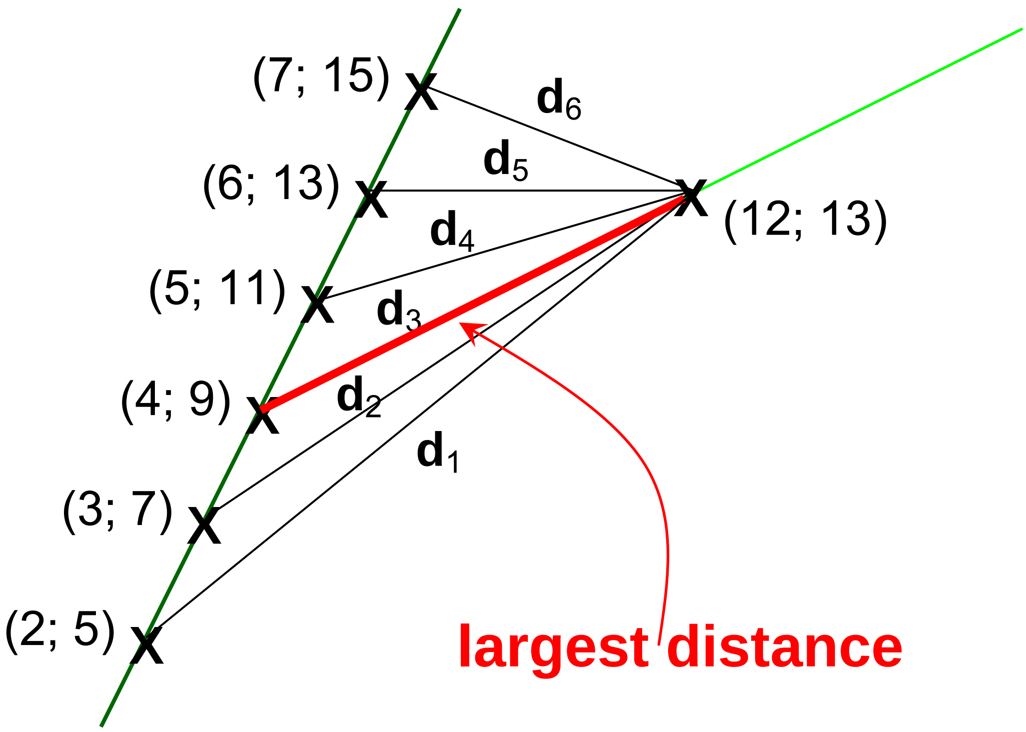
$\Rightarrow$ time-like regression line

Conclusion:

This is the **largest distance** between the space-time point (12; 13) and the green regression line.

12 space-like unit steps to the right into the direction of the x-axis

13 time-like unit steps upwards into the direction of the ct-axis




The distance vectors $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_4$, $\mathbf{d}_5$, and $\mathbf{d}_6$ are shorter than the longest distance vector $\mathbf{d}_3$.

# Orthogonal Regression in Spacetime

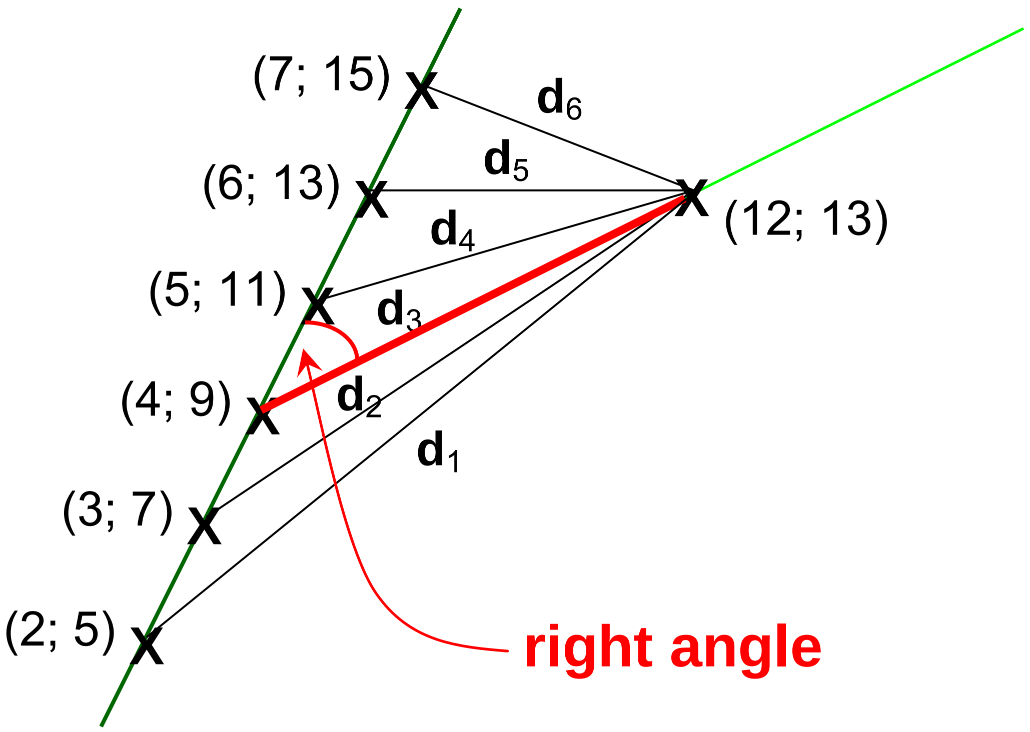
*(Orthogonale Regression in der Raumzeit)*

(1) $y = ct = 2x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line

Another Conclusion:

This is a **right angle**! The green regression line and the largest distance vector $\mathbf{d}_3$ are perpendicular to each other.



# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

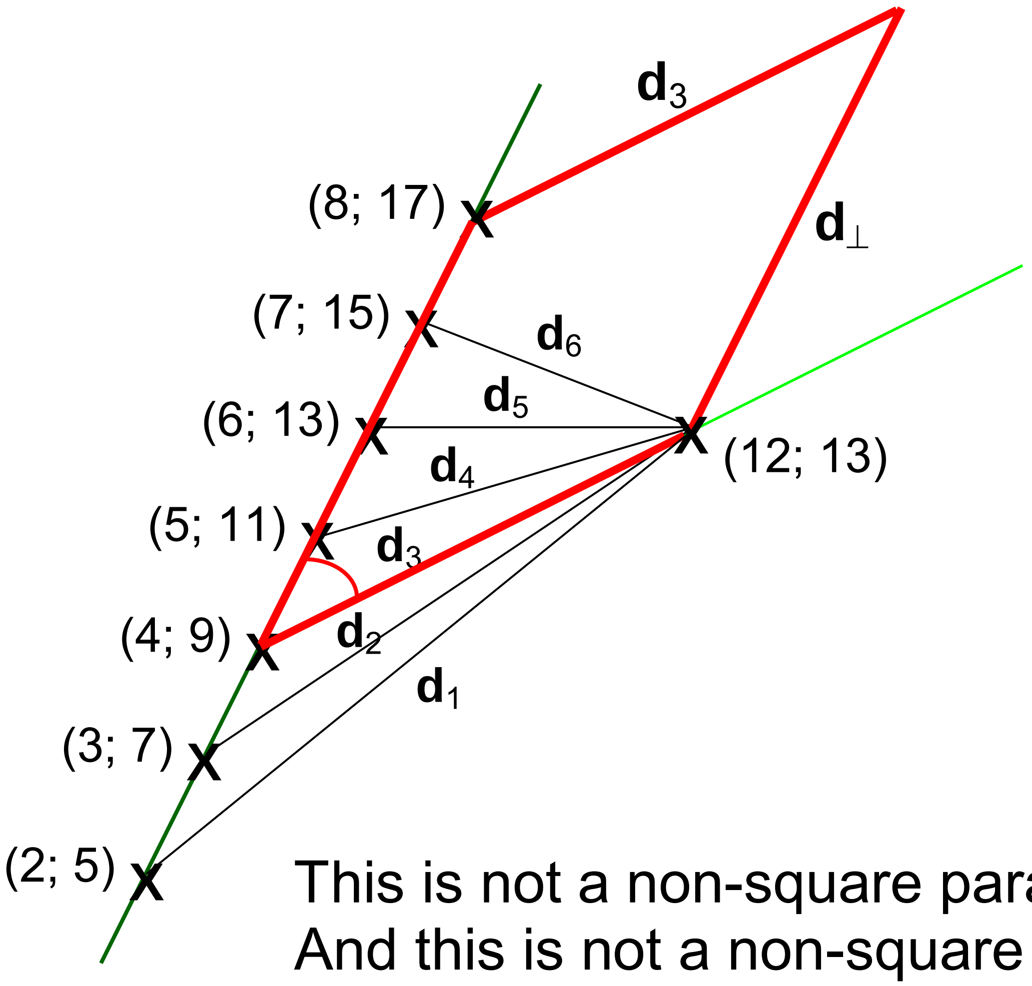
(1) $y = c\,t = 2\,x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line

Next Conclusion:

And this is a **spacetime square**:

$\mathbf{d}_3 = 8\,\gamma_x + 4\,\gamma_t \quad \Rightarrow \quad \mathbf{d}_\perp = 4\,\gamma_x + 8\,\gamma_t$




This is not a non-square parallelogram!
And this is not a non-square rhombus!
In spacetime this is a square!

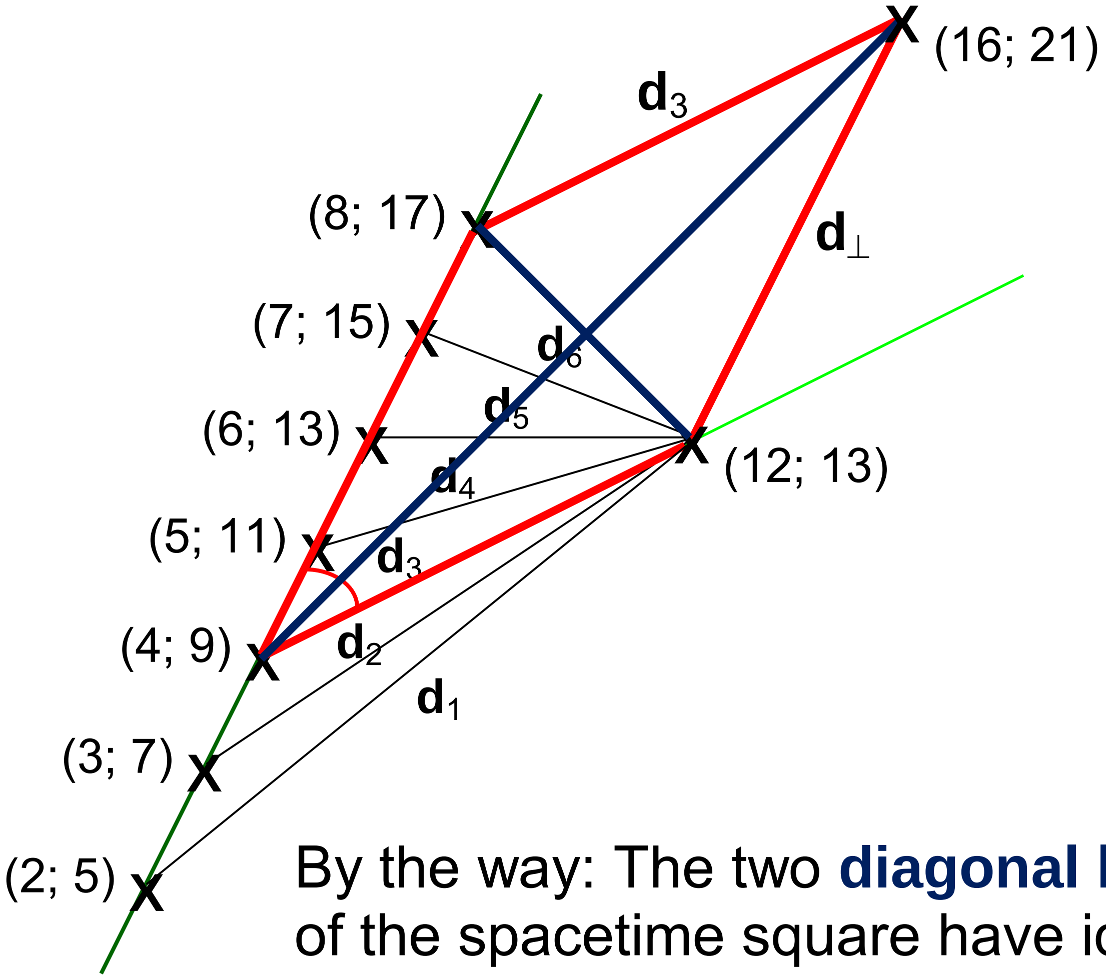
# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(1) $y = ct = 2x + 1 \quad \Rightarrow \quad m^2 > 1$

$\Rightarrow$ time-like regression line

And this is a **spacetime square**:




By the way: The two **diagonal lines** of the spacetime square have identical lengths: Their length is equal to zero.

First diagonal line:

$((16 - 4)\,\gamma_x + (21 - 9)\,\gamma_t)^2 = (12\,\gamma_x + 12\,\gamma_t)^2 = -144 + 144 = 0$

Second diagonal line:

$((12 - 8)\,\gamma_x + (13 - 17)\,\gamma_t)^2 = (4\,\gamma_x - 4\,\gamma_t)^2 = -16 + 16 = 0$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Finally we will check the equation of $x^+$ we have found

$$x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

with the given values of this first example (1):

measured data point (12; 13) $\Rightarrow$ $x = 12$, $y = c\,t = 13$

regression line: $y = c\,t = 2\,x + 1$ $\Rightarrow$ $m = 2$, $b = 1$

$$\Rightarrow \quad x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

$$= \frac{12 - 2 \cdot 13 + 2 \cdot 1}{1 - 2^2}$$

$$= \frac{-12}{-3}$$

$$= 4$$

And this is indeed the value of the x-coordinate of the estimated point (4; 9) on the regression line with the largest distance to the data point.

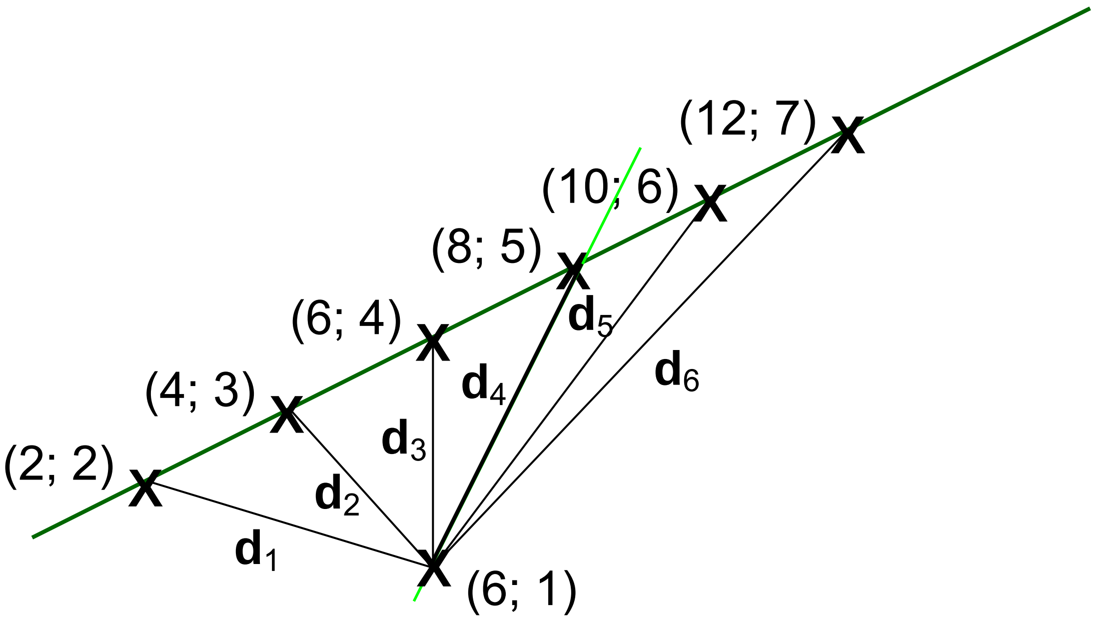
# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Some Examples:

(2) $y = c\,t = 0.5\,x + 1 \Rightarrow \; m^2 < 1$

$\Rightarrow$ space-like regression line




$\mathbf{d}_6 = (6 - 12)\,\gamma_x + (1 - 7)\,\gamma_t = -6\,\gamma_x - 6\,\gamma_t$

$\mathbf{d}_5 = (6 - 10)\,\gamma_x + (1 - 6)\,\gamma_t = -4\,\gamma_x - 5\,\gamma_t$

$\mathbf{d}_4 = (6 - 8)\,\gamma_x + (1 - 5)\,\gamma_t = -2\,\gamma_x - 4\,\gamma_t$

$\mathbf{d}_3 = (6 - 6)\,\gamma_x + (1 - 4)\,\gamma_t = -3\,\gamma_t$

$\mathbf{d}_2 = (6 - 4)\,\gamma_x + (1 - 3)\,\gamma_t = 2\,\gamma_x - 2\,\gamma_t$

$\mathbf{d}_1 = (6 - 2)\,\gamma_x + (1 - 2)\,\gamma_t = 4\,\gamma_x - 1\,\gamma_t$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(2) $y = c\,t = 0.5\,x + 1 \Rightarrow m^2 < 1$

$\Rightarrow$ space-like regression line

$\mathbf{d}_6 = (6 - 12)\,\gamma_x + (1 - 7)\,\gamma_t = -6\,\gamma_x - 6\,\gamma_t$

$\mathbf{d}_5 = (6 - 10)\,\gamma_x + (1 - 6)\,\gamma_t = -4\,\gamma_x - 5\,\gamma_t$

$\mathbf{d}_4 = (6 - 8)\,\gamma_x + (1 - 5)\,\gamma_t = -2\,\gamma_x - 4\,\gamma_t$

$\mathbf{d}_3 = (6 - 6)\,\gamma_x + (1 - 4)\,\gamma_t = -3\,\gamma_t$

$\mathbf{d}_2 = (6 - 4)\,\gamma_x + (1 - 3)\,\gamma_t = 2\,\gamma_x - 2\,\gamma_t$

$\mathbf{d}_1 = (6 - 2)\,\gamma_x + (1 - 2)\,\gamma_t = 4\,\gamma_x - 1\,\gamma_t$

Squares of the distance vectors $\mathbf{d}_i$:

$\mathbf{d}_6^2 = (-6\,\gamma_x - 6\,\gamma_t)^2 = -36 + 36 = 0$

$\mathbf{d}_5^2 = (-4\,\gamma_x - 5\,\gamma_t)^2 = -16 + 25 = 9$

$\mathbf{d}_4^2 = (-2\,\gamma_x - 4\,\gamma_t)^2 = -4 + 16 = 12$

$\mathbf{d}_3^2 = (-3\,\gamma_t)^2 = 9$

$\mathbf{d}_2^2 = (2\,\gamma_x - 2\,\gamma_t)^2 = -4 + 4 = 0$

$\mathbf{d}_1^2 = (4\,\gamma_x - 1\,\gamma_t)^2 = -16 + 1 = -15$

Maximum of squares

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(2) $y = ct = 0.5\,x + 1 \Rightarrow m^2 < 1$

$\Rightarrow$ space-like regression line

The magnitude of the distance vectors $\mathbf{d}_i$ is identical to the length of the distance vectors:

$$\mathbf{d}_6^2 = 0 \Rightarrow |\mathbf{d}_6^2| = 0 \Rightarrow |\mathbf{d}_6| = 0$$

$$\mathbf{d}_5^2 = 9 \Rightarrow |\mathbf{d}_5^2| = 9 \Rightarrow |\mathbf{d}_5| = 3$$

$$\mathbf{d}_4^2 = 12 \Rightarrow |\mathbf{d}_4^2| = 12 \Rightarrow |\mathbf{d}_4| = 2\sqrt{3} \approx 3.46$$

$$\mathbf{d}_3^2 = 9 \Rightarrow |\mathbf{d}_3^2| = 9 \Rightarrow |\mathbf{d}_3| = 3$$

$$\mathbf{d}_2^2 = 0 \Rightarrow |\mathbf{d}_2^2| = 0 \Rightarrow |\mathbf{d}_2| = 0$$

$$\mathbf{d}_1^2 = -15 \Rightarrow \text{no time-like vector}$$

The maximum of squares is identical to the maximum of the lengths.

$\Rightarrow$ The squares are maximized now.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(2) $y = c\,t = 0.5\,x + 1 \Rightarrow \; m^2 < 1$

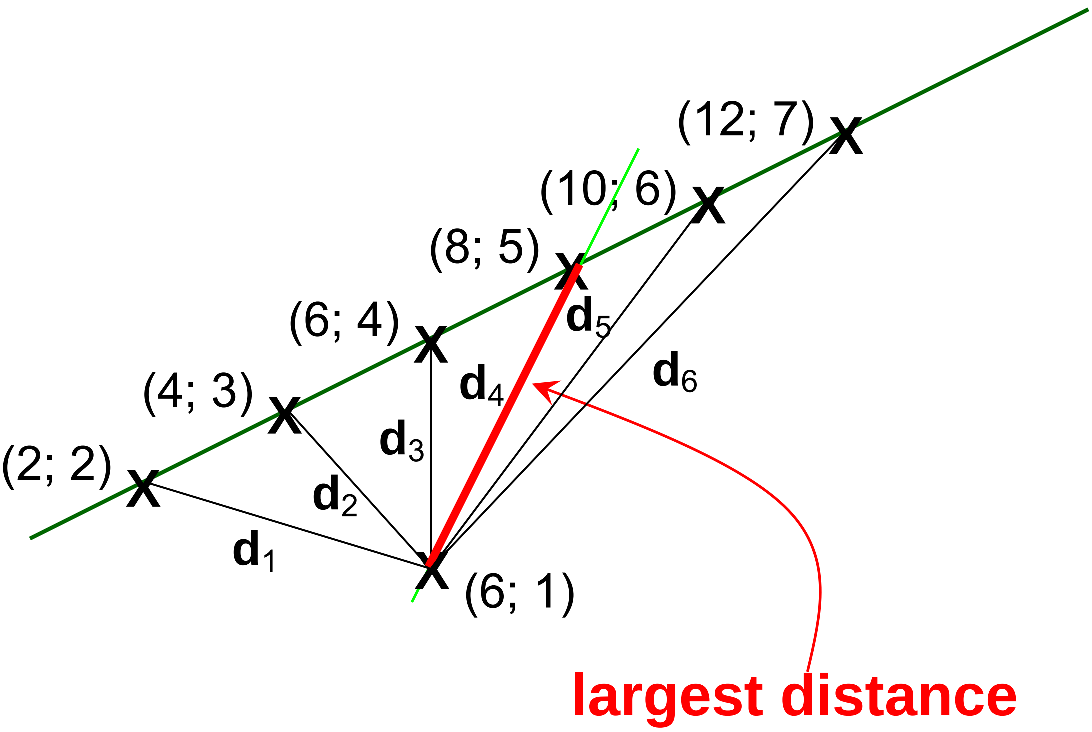
$\Rightarrow$ space-like regression line

Conclusion:

This is the **largest distance** between the space-time point (6; 1) and the green regression line.

6 space-like unit steps to the right into the direction of the x-axis

1 time-like unit step upwards into the direction of the ct-axis




The distance vectors $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, $\mathbf{d}_5$, and $\mathbf{d}_6$ are shorter than the longest distance vector $\mathbf{d}_4$.

# Orthogonal Regression in Spacetime

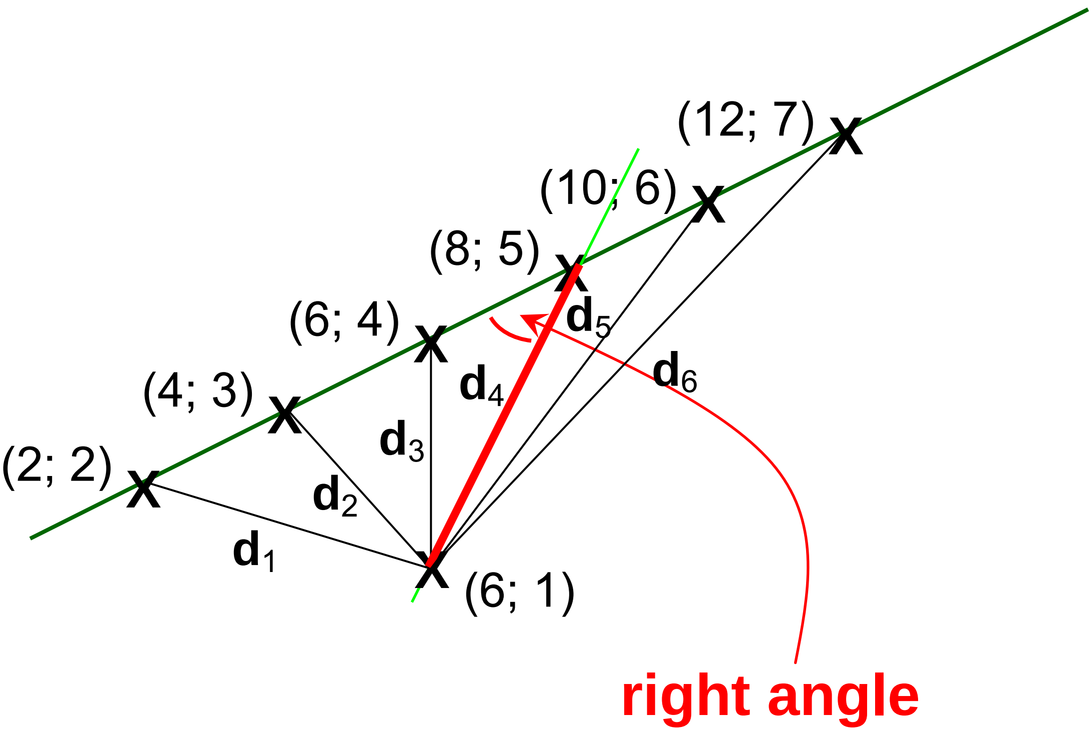
*(Orthogonale Regression in der Raumzeit)*

(2) $y = c\,t = 0.5\,x + 1 \Rightarrow \; m^2 < 1$

$\Rightarrow$ space-like regression line

Another Conclusion:

This is a **right angle**! The green regression line and the largest distance vector $\mathbf{d}_4$ are perpendicular to each other.



# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

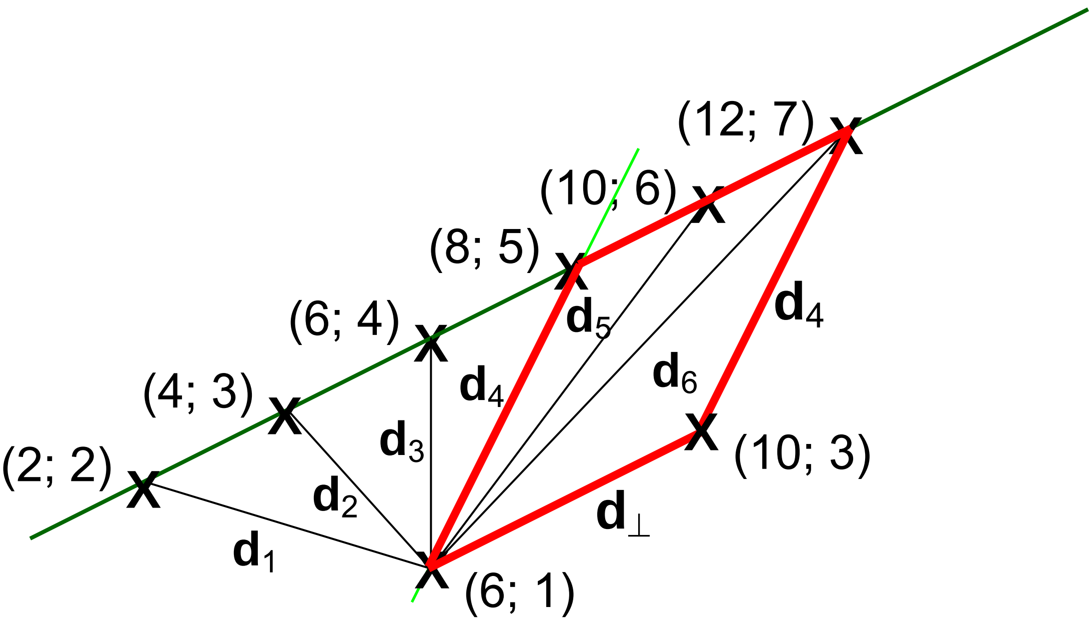
(2) $y = c\,t = 0.5\,x + 1 \Rightarrow \; m^2 < 1$

$\Rightarrow$ space-like regression line

Next Conclusion:

And this is a **spacetime square**:

$\mathbf{d}_4 = -\,2\,\gamma_x - 4\,\gamma_t \qquad \Rightarrow \qquad \mathbf{d}_\perp = 4\,\gamma_x + 2\,\gamma_t$




This is not a non-square parallelogram!
And this is not a non-square rhombus!
In spacetime this is a square!

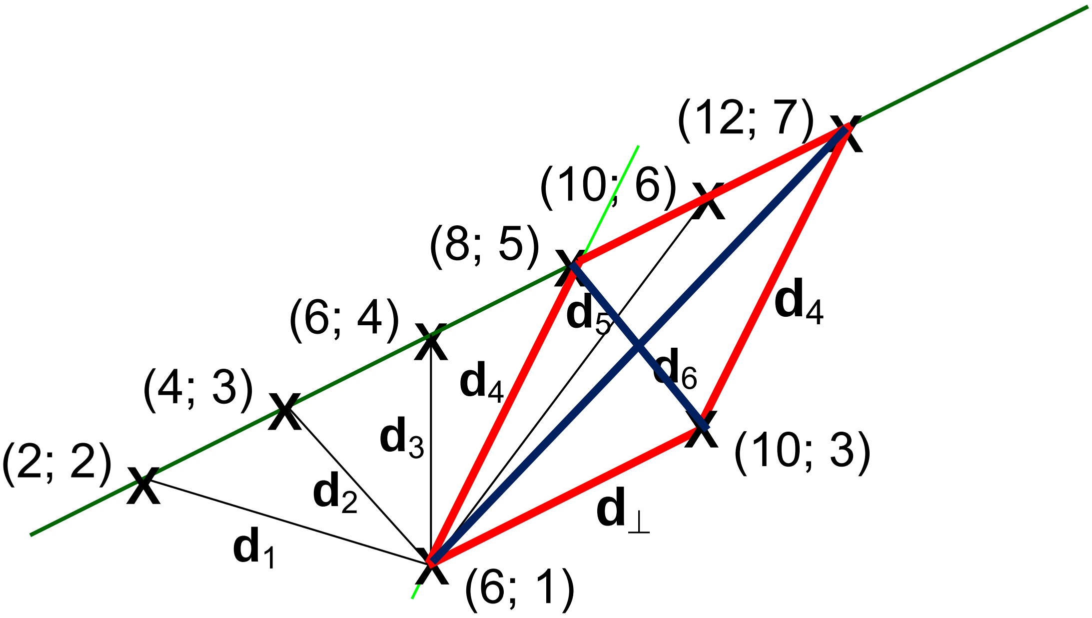
# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

(2) $y = ct = 0.5\,x + 1 \Rightarrow m^2 < 1$

$\Rightarrow$ space-like regression line

And this is a **spacetime square**:




By the way: The two **diagonal lines** of the spacetime square have identical lengths: Their length is equal to zero.

First diagonal line:

$((6 - 12)\,\gamma_x + (1 - 7)\,\gamma_t)^2 = (-6\,\gamma_x - 6\,\gamma_t)^2 = -36 + 36 = 0$

Second diagonal line:

$((10 - 8)\,\gamma_x + (3 - 5)\,\gamma_t)^2 = (2\,\gamma_x - 2\,\gamma_t)^2 = -4 + 4 = 0$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Finally we will check the equation of $x^+$ we have found

$$x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

with the given values of this second example (2):

measured data point (6; 1) $\Rightarrow$ $x = 6$, $y = c\,t = 1$

regression line: $y = c\,t = 0.5\,x + 1$ $\Rightarrow$ $m = 0.5$, $b = 1$

$$\Rightarrow \quad x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

$$= \frac{6 - 0.5 \cdot 1 + 0.5 \cdot 1}{1 - 0.5^2}$$

$$= \frac{6}{0.75}$$

$$= 8$$

And this is indeed the value of the x-coordinate of the estimated point (8; 5) on the regression line with the largest distance to the data point.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Now we will go on to find the best-fit line of orthogonal regression in spacetime and substitute the value

$$x^+ = \frac{x - m\,y + m\,b}{1 - m^2}$$

of maximized squared spacetime distances into the equation of the square of the height:

$$h^2 = -\,(x - x^+)^2 + (y - y^+)^2$$

$$= -\,(x - x^+)^2 + (y - m\,x^+ - b)^2$$

$$= -\left(x - \frac{x - m\,y + m\,b}{1 - m^2}\right)^2 + \left(y - m\,\frac{x - m\,y + m\,b}{1 - m^2} - b\right)^2$$

$$= -\left(x + \frac{-x + m\,y - m\,b}{1 - m^2}\right)^2 + \left(y + \frac{-m\,x + m^2\,y - m^2\,b}{1 - m^2} - b\right)^2$$

$$= -\left(\frac{m}{1 - m^2}\,(y - m\,x - b)\right)^2 + \left(\frac{1}{1 - m^2}\,(y - m\,x - b)\right)^2$$

$$= \left(1 - m^2\right)\left(\frac{y - m\,x - b}{1 - m^2}\right)^2$$

$$= \frac{(y - m\,x - b)^2}{1 - m^2} = \frac{{e_y}^2}{1 - m^2}$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

spacetime:

$$\Rightarrow\; h^2 = -\,(x - x^+)^2 + (y - y^+)^2$$

$$= -\,(x - x^+)^2 + (y - m\,x^+ - b)^2$$

$$= \frac{(y - m\,x - b)^2}{1 - m^2} = \frac{{e_y}^2}{1 - m^2}$$

This result is nearly identical to the result, we have reached earlier when looking for the sum of orthogonal squared errors of conventional orthogonal regression:

*pure space*:

$$h^2 = (x - x^*)^2 + (y - y^*)^2$$

$$= (x - x^*)^2 + (y - m\,x^* - b)^2$$

$$= \frac{(y - m\,x - b)^2}{1 + m^2} = \frac{{e_y}^2}{1 + m^2}$$

The only formal difference is the negative sign of the slope of the regression line we want to find.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Sum of squared orthogonal errors in spacetime:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1 - m^2}$$

Partial derivative of the sum of squared orthogonal errors in spacetime with respect to the variable b (y-intercept or ct-intercept):

$$\frac{\partial\,\text{SSOE}}{\partial b} = \frac{\partial}{\partial b}\left( \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1 - m^2} \right)$$

$$= \sum_{i=1}^{n} \left( \frac{\partial}{\partial b} \frac{(y_i - m\,x_i - b)^2}{1 - m^2} \right)$$

$$= \sum_{i=1}^{n} \frac{2\,(y_i - m\,x_i - b)\,(-1)}{1 - m^2}$$

$$= -2 \sum_{i=1}^{n} \frac{y_i - m\,x_i - b}{1 - m^2}$$

$$= -\frac{2}{1 - m^2} \sum_{i=1}^{n} (y_i - m\,x_i - b)$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Sum of squared orthogonal errors:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1 - m^2}$$

Partial derivative of the sum of squared errors with respect to the variable b:

$$\frac{\partial\,\text{SSOE}}{\partial b} = -\frac{2}{1 - m^2} \sum_{i=1}^{n} (y_i - m\,x_i - b)$$

Stationary values:

$$\frac{\partial\,\text{SSOE}}{\partial b} = -\frac{2}{1 - m^2} \sum_{i=1}^{n} (y_i - m\,x_i - b) = 0$$

$$\Rightarrow \qquad \sum_{i=1}^{n} (y_i - m\,x_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + \ldots + 1}_{\text{n times 1}} = n$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Stationary values:

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \quad b\, n \quad = 0$$

Now we divide by n:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - m\, \frac{1}{n}\sum_{i=1}^{n} x_i - b \quad = 0$$

And substituting …

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$ ………… arithmetic mean of y

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$ ………… arithmetic mean of x

… will result in:

$$\bar{y} - m\,\bar{x} - b \;= 0$$

$\Rightarrow$

$$\boxed{\bar{y} = m\,\bar{x} + b}$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Sum of squared orthogonal errors:

$$\text{SSOE} = \sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1 - m^2}$$

Now the partial derivative of the sum of squared orthogonal errors in spacetime will be found with respect to the variable m (slope):

$$\frac{\partial\,\text{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n} \frac{(y_i - m\,x_i - b)^2}{1 - m^2}\right)$$

$$= \sum_{i=1}^{n}\left(\frac{\partial}{\partial m}\,\frac{(y_i - m\,x_i - b)^2}{1 - m^2}\right)$$

$$= \sum_{i=1}^{n}\frac{2\,(1-m^2)\,(y_i - m\,x_i - b)\,(-x_i) - (y_i - m\,x_i - b)^2\,(-2\,m)}{(1-m^2)^2}$$

$$= \frac{2}{(1-m^2)^2}\sum_{i=1}^{n}\Big((-x_i + m^2\,x_i)\,(y_i - m\,x_i - b) + m\,(y_i - m\,x_i - b)^2\Big)$$

$$= \frac{2}{(1-m^2)^2}\sum_{i=1}^{n}\Big((-x_i + m^2\,x_i + m\,y_i - m^2\,x_i - m\,b)\,(y_i - m\,x_i - b)\Big)$$

$$= \frac{2}{(1-m^2)^2}\sum_{i=1}^{n}\Big((-x_i + m\,y_i - m\,b)\,(y_i - m\,x_i - b)\Big)$$

$$= \frac{2}{(1-m^2)^2} \sum_{i=1}^{n} (-x_i y_i + m x_i^2 + b x_i + m y_i^2 - m^2 x_i y_i - mb y_i$$
$$- mb y_i + m^2 b x_i + mb^2)$$

$$= \frac{2}{(1-m^2)^2} \sum_{i=1}^{n} (-x_i y_i + m x_i^2 + b x_i + m y_i^2 - m^2 x_i y_i$$
$$- 2mb y_i + m^2 b x_i + mb^2)$$

$$= \frac{2}{(1-m^2)^2} \sum_{i=1}^{n} ((1+m^2) b x_i - 2mb y_i - (1+m^2) x_i y_i$$
$$+ m x_i^2 + m y_i^2 + mb^2)$$

$$= \frac{2b(1+m^2)}{(1-m^2)^2} \sum_{i=1}^{n} x_i - \frac{4mb}{(1-m^2)^2} \sum_{i=1}^{n} y_i - \frac{2(1+m^2)}{(1-m^2)^2} \sum_{i=1}^{n} x_i y_i$$
$$+ \frac{2m}{(1-m^2)^2} \sum_{i=1}^{n} x_i^2 + \frac{2m}{(1-m^2)^2} \sum_{i=1}^{n} y_i^2 + \frac{2mb^2}{(1-m^2)^2} \sum_{i=1}^{n} 1$$

$$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + 1 + \ldots + 1}_{\text{n times 1}} = n$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

$$\frac{\partial\,\mathrm{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n}\frac{(y_i - m\,x_i - b)^2}{1-m^2}\right)$$

$$= \frac{2\,b\,(1+m^2)}{(1-m^2)^2}\sum_{i=1}^{n}x_i - \frac{4\,m\,b}{(1-m^2)^2}\sum_{i=1}^{n}y_i - \frac{2\,(1+m^2)}{(1-m^2)^2}\sum_{i=1}^{n}x_i\,y_i$$
$$+ \frac{2\,m}{(1-m^2)^2}\sum_{i=1}^{n}x_i^{\,2} + \frac{2\,m}{(1-m^2)^2}\sum_{i=1}^{n}y_i^{\,2} + \frac{2\,m\,b^2}{(1-m^2)^2}\,n$$

Stationary values:

$$\frac{\partial\,\mathrm{SSOE}}{\partial m} = \frac{\partial}{\partial m}\left(\sum_{i=1}^{n}\frac{(y_i - m\,x_i - b)^2}{1-m^2}\right) = 0$$

$$0 = \frac{2\,b\,(1+m^2)}{(1-m^2)^2}\sum_{i=1}^{n}x_i - \frac{4\,m\,b}{(1-m^2)^2}\sum_{i=1}^{n}y_i - \frac{2\,(1+m^2)}{(1-m^2)^2}\sum_{i=1}^{n}x_i\,y_i$$
$$+ \frac{2\,m}{(1-m^2)^2}\sum_{i=1}^{n}x_i^{\,2} + \frac{2\,m}{(1-m^2)^2}\sum_{i=1}^{n}y_i^{\,2} + \frac{2\,m\,b^2}{(1-m^2)^2}\,n$$

Now we divide again by n …

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Now we divide again by n and multiply by the denominator $(1 - m^2)^2$ …

$$0 = 2\,b\,(1 + m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i - 4\,m\,b\cdot\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$- 2\,(1 + m^2)\cdot\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i + 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$$

$$+ 2\,m\cdot\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} + 2\,m\,b^2$$

… and substitute:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} = s_x^{\,2} + \bar{x}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} = s_y^{\,2} + \bar{y}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i = s_{xy} + \bar{x}\,\bar{y}$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

$$0 = 2\,b\,(1 + m^2) \cdot \frac{1}{n}\sum_{i=1}^{n} x_i - 4\,m\,b \cdot \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$- 2\,(1 + m^2) \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\,y_i + 2\,m \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$+ 2\,m \cdot \frac{1}{n}\sum_{i=1}^{n} y_i^2 + 2\,m\,b^2$$

$$0 = 2\,b\,(1 + m^2)\,\bar{x} - 4\,m\,b\,\bar{y} - 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^2 + \bar{x}^2) + 2\,m\,(s_y^2 + \bar{y}^2) + 2\,m\,b^2$$

Together with the previous result

$$\bar{y} = m\,\bar{x} + b$$

it should be possible to solve these two equations for m and b.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

$$0 = 2\,b\,(1 + m^2)\,\bar{x} - 4\,m\,b\,\bar{y} - 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^2 + \bar{x}^2) + 2\,m\,(s_y^2 + \bar{y}^2) + 2\,m\,b^2$$

$$0 = 2\,b\,(1 + m^2)\,\bar{x} - 4\,m^2\,b\,\bar{x} - 4\,m\,b^2$$

$$- 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^2 + \bar{x}^2) + 2\,m\,(s_y^2 + \bar{y}^2) + 2\,m\,b^2$$

$$0 = 2\,b\,(1 - m^2)\,\bar{x} - 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^2 + \bar{x}^2) + 2\,m\,(s_y^2 + \bar{y}^2) - 2\,m\,b^2$$

$$\bar{y} - m\,\bar{x} = b \qquad\qquad \bar{y}^2 - 2\,m\,\bar{x}\,\bar{y} + m^2\,\bar{x}^2 = b^2$$

$$0 = 2\,(1 - m^2)\,\bar{x}\,\bar{y} - 2\,m\,(1 - m^2)\,\bar{x}^2$$

$$- 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$

$$+ 2\,m\,(s_x^2 + \bar{x}^2) + 2\,m\,(s_y^2 + \bar{y}^2)$$

$$- 2\,m\,\bar{y}^2 + 4\,m^2\,\bar{x}\,\bar{y} - 2\,m^3\,\bar{x}^2$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

$$0 = 2\,(1 - m^2)\,\bar{x}\,\bar{y} - 2\,m\,(1 - m^2)\,\bar{x}^2$$
$$- 2\,(1 + m^2)\,(s_{xy} + \bar{x}\,\bar{y})$$
$$+ 2\,m\,(s_x^{\;2} + \bar{x}^2) + 2\,m\,(s_y^{\;2} + \bar{y}^2)$$
$$- 2\,m\,\bar{y}^2 + 4\,m^2\,\bar{x}\,\bar{y} - 2\,m^3\,\bar{x}^2$$

Some terms can be cancelled …

$$0 = -4\,m^2\,\bar{x}\,\bar{y} - 2\,(1 + m^2)\,s_{xy} + 2\,m\,s_x^{\;2}$$
$$+ 2\,m\,(s_y^{\;2} + \bar{y}^2) - 2\,m\,\bar{y}^2 + 4\,m^2\,\bar{x}\,\bar{y}$$

… and cancelled …

$$0 = -2\,(1 + m^2)\,s_{xy} + 2\,m\,s_x^{\;2} + 2\,m\,s_y^{\;2}$$

$$\Rightarrow\; 0 = (1 + m^2)\,s_{xy} - m\,s_x^{\;2} - m\,s_y^{\;2}$$

This looks like a quadratic equation:

$$0 = s_{xy}\,m^2 - (s_x^{\;2} + s_y^{\;2})\,m + s_{xy}$$

$$\Rightarrow\; 0 = m^2 - \frac{s_x^{\;2} + s_y^{\;2}}{s_{xy}}\,m + 1$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

After completing the square,

$$0 = m^2 - \frac{s_x^2 + s_y^2}{s_{xy}} m + 1$$

$$0 = m^2 - \frac{s_x^2 + s_y^2}{s_{xy}} m + \frac{(s_x^2 + s_y^2)^2}{4 s_{xy}^2} - \frac{(s_x^2 + s_y^2)^2}{4 s_{xy}^2} + 1$$

$$0 = \left( m - \frac{s_x^2 + s_y^2}{2 s_{xy}} \right)^2 - \frac{(s_x^2 + s_y^2)^2}{4 s_{xy}^2} + 1$$

we get:

$$\left( m - \frac{s_x^2 + s_y^2}{2 s_{xy}} \right)^2 = \frac{(s_x^2 + s_y^2)^2}{4 s_{xy}^2} - 1$$

$$= \frac{(s_x^2 + s_y^2)^2 - 4 s_{xy}^2}{4 s_{xy}^2}$$

$$m - \frac{s_x^2 + s_y^2}{2 s_{xy}} = \pm \frac{\sqrt{(s_x^2 + s_y^2)^2 - 4 s_{xy}^2}}{2 s_{xy}}$$

$$m = \frac{s_x^2 + s_y^2 \pm \sqrt{(s_x^2 + s_y^2)^2 - 4 s_{xy}^2}}{2 s_{xy}}$$

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Intermediate preliminary result:

The spacetime regression line

$$y = c\,t = m\,x + b$$

has the slope $$m = \frac{s_x{}^2 + s_y{}^2 \pm \sqrt{(s_x{}^2 + s_y{}^2)^2 - 4\,s_{xy}{}^2}}{2\,s_{xy}}$$

and the y-intercept

$$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_x{}^2 + s_y{}^2 \pm \sqrt{(s_x{}^2 + s_y{}^2)^2 - 4\,s_{xy}{}^2}}{2\,s_{xy}}\,\bar{x}$$

when the orthogonal errors are maximized

The parameters m and b can now be called "spacetime regression coefficients".

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Again we have to decide for the correct sign, e.g. by taking the two data points (0; 0) and (1; 1) to get a slope of m = 1:

$\bar{x} = 0.5 \qquad \bar{y} = 0.5$

$s_x^2 = 0.25 \qquad s_y^2 = 0.25 \qquad s_{xy} = 0.25$

$$m = \frac{s_x^2 + s_y^2 \pm \sqrt{(s_x^2 + s_y^2)^2 - 4\, s_{xy}^2}}{2\, s_{xy}}$$

$$= \frac{0.25 + 0.25 \pm \sqrt{(0.25 + 0.25)^2 - 4 \cdot 0.25^2}}{2 \cdot 0.25}$$

$$= \frac{2 \cdot 0.25 \pm \sqrt{0}}{2 \cdot 0.25} = 1$$

$\Rightarrow$ The sign does not matter, as we have chosen a light-like regression line with zero length.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Next try: We want to get a time-like regression line and therefore take the two data points (0; 0) and (1; 2) to get a slope of m = 2:

$\bar{x} = 0.5 \qquad \bar{y} = 1$

$s_x^2 = 0.25 \qquad s_y^2 = 1 \qquad s_{xy} = 0.5$

$$m = \frac{s_x^2 + s_y^2 \pm \sqrt{(s_x^2 + s_y^2)^2 - 4\, s_{xy}^2}}{2\, s_{xy}}$$

$$= \frac{0.25 + 1 \pm \sqrt{(0.25 + 1)^2 - 4 \cdot 0.5^2}}{2 \cdot 0.5}$$

$$= \frac{1.25 \pm \sqrt{0.5625}}{1}$$

$= 1.25 \pm 0.75 \quad \Rightarrow \quad m = 1.25 + 0.75 = 2$
if the positive sign is chosen.

$\Rightarrow$ The sign has to be positive if the orthogonal regression line is a time-like line.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

And another next try: We want to get a space-like regression line and therefore take the two data points (0; 0) and (2; 1) to get a slope of m = 0.5:

$\bar{x} = 1 \qquad \bar{y} = 0.5$

$s_x^2 = 1 \qquad s_y^2 = 0.25 \qquad s_{xy} = 0.5$

$$m = \frac{s_x^2 + s_y^2 \pm \sqrt{(s_x^2 + s_y^2)^2 - 4\, s_{xy}^2}}{2\, s_{xy}}$$

$$= \frac{1 + 0.25 \pm \sqrt{(1 + 0.25)^2 - 4 \cdot 0.5^2}}{2 \cdot 0.5}$$

$$= \frac{1.25 \pm \sqrt{0.5625}}{1}$$

$= 1.25 \pm 0.75 \quad \Rightarrow \quad m = 1.25 – 0.75 = 0.5$ if the negative sign is chosen.

$\Rightarrow$ The sign has to be negative if the orthogonal regression line is a space-like line.

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

First final result:

The spacetime regression line

$$y = c\,t = m\,x + b$$

has the slope $$m = \frac{s_x^{\ 2} + s_y^{\ 2} + \sqrt{(s_x^{\ 2} + s_y^{\ 2})^2 - 4\,s_{xy}^{\ 2}}}{2\,s_{xy}}$$

and the y-intercept

$$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_x^{\ 2} + s_y^{\ 2} + \sqrt{(s_x^{\ 2} + s_y^{\ 2})^2 - 4\,s_{xy}^{\ 2}}}{2\,s_{xy}}\,\bar{x}$$

when the orthogonal errors are maximized to get a time-like regression line.

The parameters m and b can now be called "spacetime regression coefficients of time-like regression".

# Orthogonal Regression in Spacetime

*(Orthogonale Regression in der Raumzeit)*

Second final result:

The spacetime regression line

$$y = c\,t = m\,x + b$$

has the slope $$m = \frac{s_x{}^2 + s_y{}^2 - \sqrt{(s_x{}^2 + s_y{}^2)^2 - 4\,s_{xy}{}^2}}{2\,s_{xy}}$$

and the y-intercept

$$b = \bar{y} - m\,\bar{x} = \bar{y} - \frac{s_x{}^2 + s_y{}^2 - \sqrt{(s_x{}^2 + s_y{}^2)^2 - 4\,s_{xy}{}^2}}{2\,s_{xy}}\,\bar{x}$$

when the orthogonal errors are maximized to get a space-like regression line.

The parameters m and b can now be called “spacetime regression coefficients of space-like regression”.

# Spacetime example Problem

*(Raumzeitliche Beispielaufgabe)*

The following data points are given:

| $x_i$ | $ct_i = y_i$ |
|---|---|
| 2 | 4 |
| 3 | 5 |
| 4 | 5.5 |
| 5 | 5.5 |
| 6 | 6 |

Please find the orthogonal regression line (orthogonal best-fit line) in spacetime.

# Solution

*(Lösung)*

Table of data points:

| $x_i$ | $y_i$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|
| 2 | 4 | 4 | 1.44 | 2.4 |
| 3 | 5 | 1 | 0.04 | 0.2 |
| 4 | 5.5 | 0 | 0.09 | 0 |
| 5 | 5.5 | 1 | 0.09 | 0.3 |
| 6 | 6 | 4 | 0.64 | 1.6 |
| 20 | 26 | 10 | 2.30 | 4.5 |

$$\bar{x} = \frac{20}{5} = 4 \qquad s_x^2 = \frac{10}{5} = 2$$

$$\bar{y} = \frac{26}{5} = 5.2 \qquad s_y^2 = \frac{2.3}{5} = 0.46$$

$$s_{xy} = \frac{4.5}{5} = 0.9$$

# Solution

*(Lösung)*

Checks:

$$s_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2 = \frac{90}{5} - 4^2 = 2$$

$$s_y^{\,2} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\,2} - \bar{y}^2 = \frac{137.5}{5} - 5.2^2 = 0.46$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y} = \frac{108.5}{5} - 4 \cdot 5.2 = 0.9$$

And all this will give a space-like regression, because the measured spatial values $x_i$ are increasing more rapidly than the measured time values $y_i = c\,t_i$:

| $x_i$ | $c\,t_i = y_i$ |
|---|---|
| 2 | 4 |
| 3 | 5 |
| 4 | 5.5 |
| 5 | 5.5 |
| 6 | 6 |

huge increase

$\Delta x_{total} = 6 - 2 = 4$

small increase

$\Delta y_{total} = 6 - 4 = 2 < \Delta x_{total}$

$\Rightarrow$ The sign has to be negative because the orthogonal regression line is a space-like line.

# Solution

*(Lösung)*

⇒ Slope of the spacetime regression line:

$$m = \frac{s_x^{\,2} + s_y^{\,2} - \sqrt{(s_x^{\,2} + s_y^{\,2})^2 - 4\,s_{xy}^{\,2}}}{2\,s_{xy}}$$

$$= \frac{2 + 0.46 - \sqrt{(2 + 0.46)^2 - 4 \cdot 0.9^2}}{2 \cdot 0.9}$$

$$= \frac{2.46 - \sqrt{2.46^2 - 3.24}}{1.8}$$

$$= \frac{2.46 - \sqrt{2.8116}}{1.8} = \frac{0.7832}{1.8}$$

$$= 0.4351 \qquad \approx 0.44$$

⇒ y-intercept of the regression line:

$$b = \bar{y} - m\,\bar{x}$$

$$= 5.2 - 0.4351 \cdot 4$$

$$= 3.459517 \qquad \approx 3.46$$

# Solution

*(Lösung)*

The spacetime regression line (best-fit line) with respect to maximized squares of orthogonal spacetime errors will be:

$$y = 0.4351\ x + 3.4595$$

Comparison: The slope of this space-like spacetime regression line is smaller than the slopes of the conventional regression lines found earlier.

$$y_{linear\text{-}y} = 0.45\ x + 3.40$$

$$y_{orthogonal} = 0.46\ x + 3.36$$

$$y_{linear\text{-}x} = 0.51\ x + 3.16$$