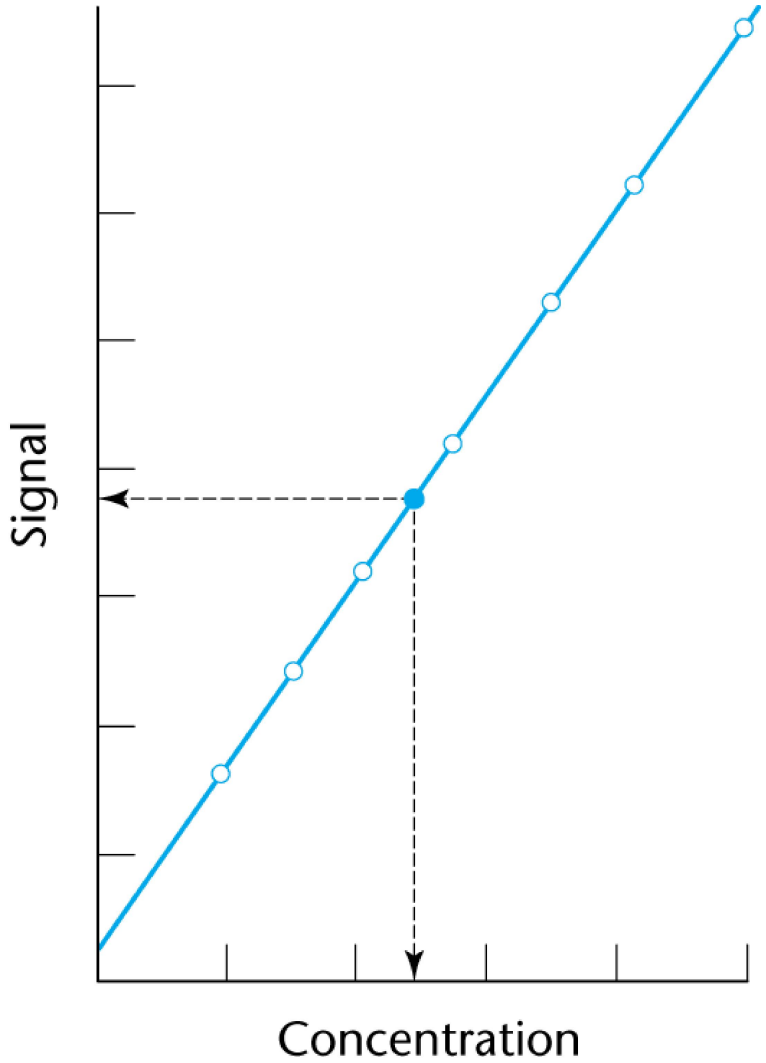
Linear Regression

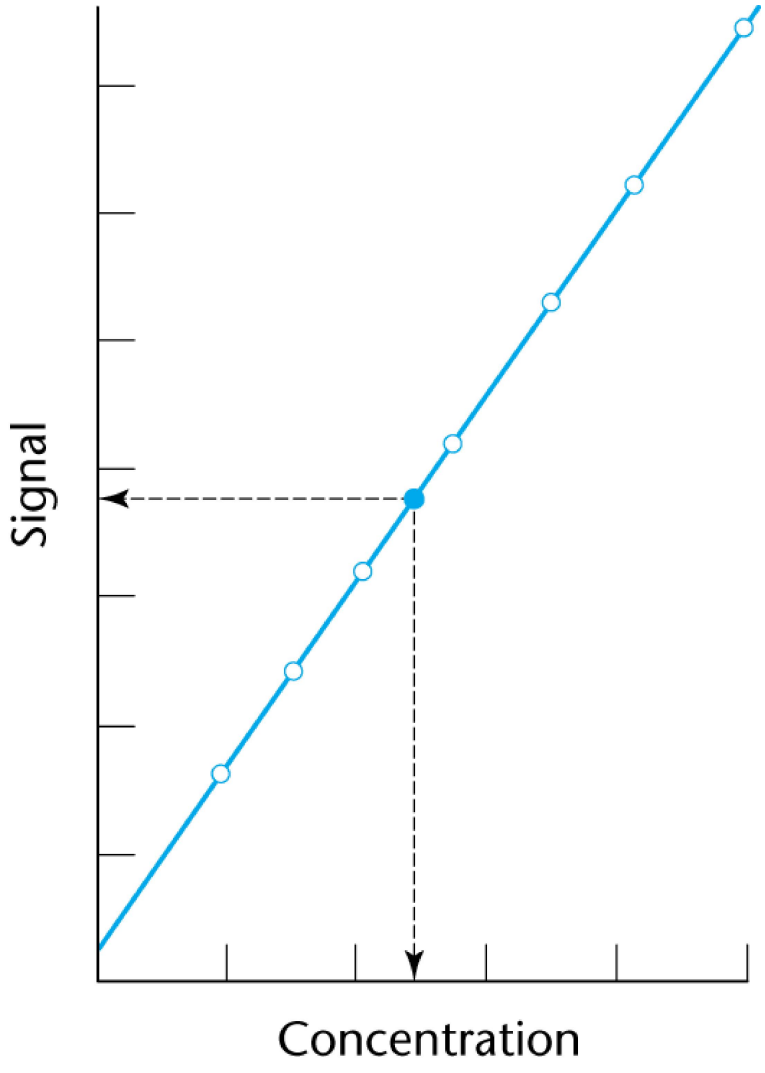
Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

November 10, 2009

Calibration



Calibration



- Take a calibaration sample with known but different concentrations.
- Based on measurements plot response curve.
- Is it linear?
- Make prediction for concentration between calibrated points.

Questions

Questions

- Is the calibration graph linear?

Questions

- Is the calibration graph linear?
- What is the best straight line fitting the data?

Questions

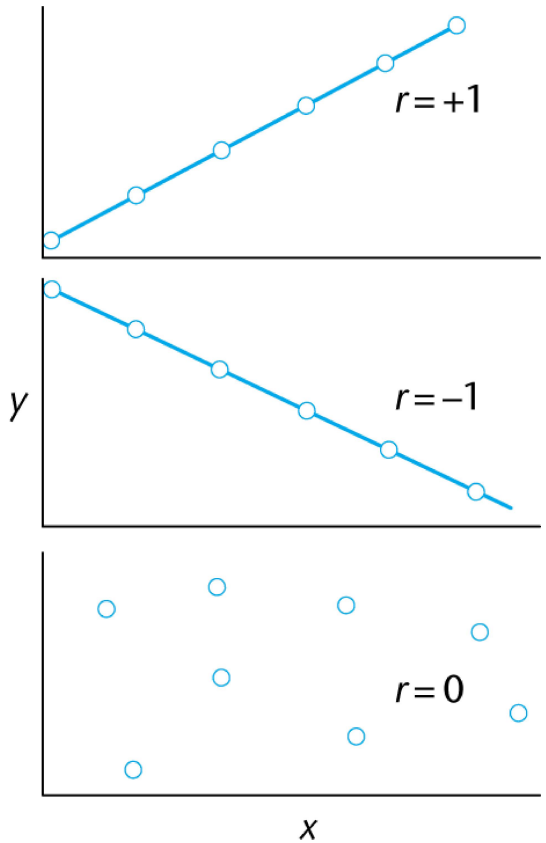
- Is the calibration graph linear?
- What is the best straight line fitting the data?
- What are the errors and confidence limits?

The correlation coefficient

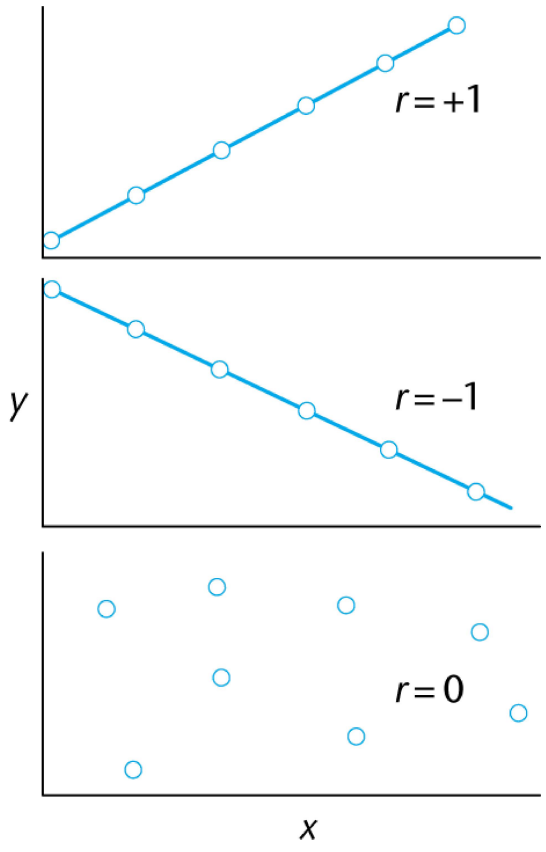
Product–moment correlation coefficient,

$$r = \frac{\sum_i \{(x_i - \bar{x})(y_i - \bar{y})\}}{\left\{\left[\sum_i (x_i - \bar{x})^2\right]\left[\sum_i (y_i - \bar{y})^2\right]\right\}^{1/2}}$$

Measure of linearity



Measure of linearity



It can be shown that the correlation coefficient satisfies

$$-1 \leq r \leq 1.$$

and $|r| \approx 1$ then the relation is close to linear.

Computations

Example 5.3.1

Standard aqueous solutions of fluorescein are examined in a fluorescence spectrometer, and yield the following fluorescence intensities (in arbitrary units):

Fluorescence intensities:	2.1	5.0	9.0	12.6	17.3	21.0	24.7
Concentration, pg ml^{-1}	0	2	4	6	8	10	12

Determine the correlation coefficient, r.

In practice, such calculations will almost certainly be performed on a calculator or computer, alongside other calculations covered below, but it is important

and instructive to examine a manually calculated result. The data are presented in a table, as follows:

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	0	2.1	−6	36	−11.0	121.00	66.0
	2	5.0	−4	16	−8.1	65.61	32.4
	4	9.0	−2	4	−4.1	16.81	8.2
	6	12.6	0	0	−0.5	0.25	0
	8	17.3	2	4	4.2	17.64	8.4
	10	21.0	4	16	7.9	62.41	31.6
	12	24.7	6	36	11.6	134.56	69.6
Sums:	42	91.7	0	112	0	418.28	216.2

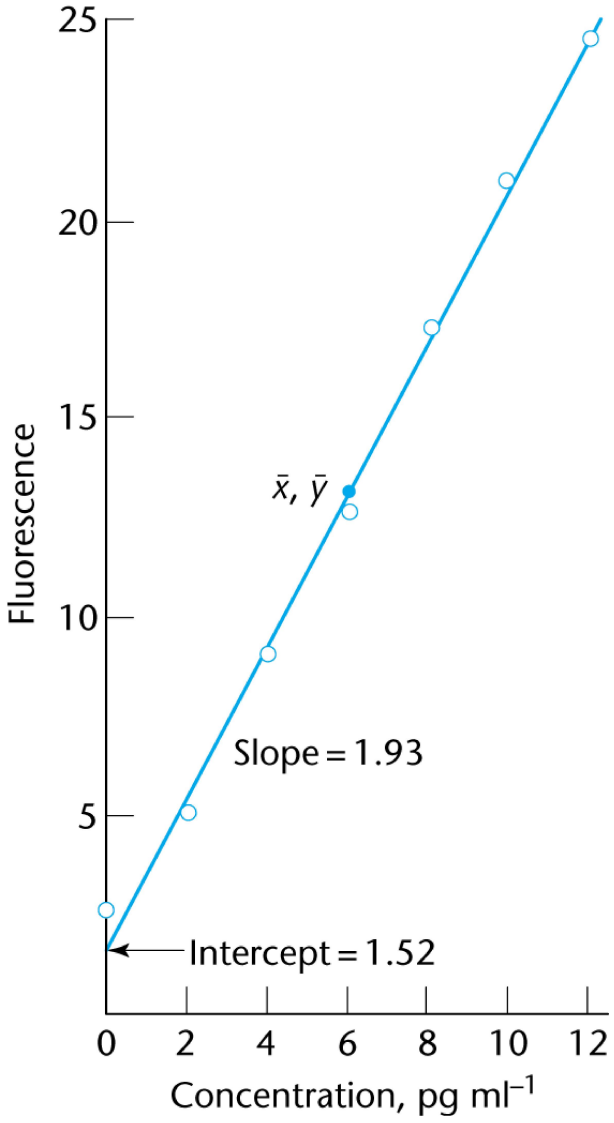
The figures below the line at the foot of the columns are in each case the su of the figures in the table: note that $\Sigma(x_i - \bar{x})$ and $\Sigma(y_i - \bar{y})$ are both zero. Us these totals in conjunction with equation (5.2), we have:

$$r = \frac{216.2}{\sqrt{112 \times 418.28}} = \frac{216.2}{216.44} = 0.9989$$

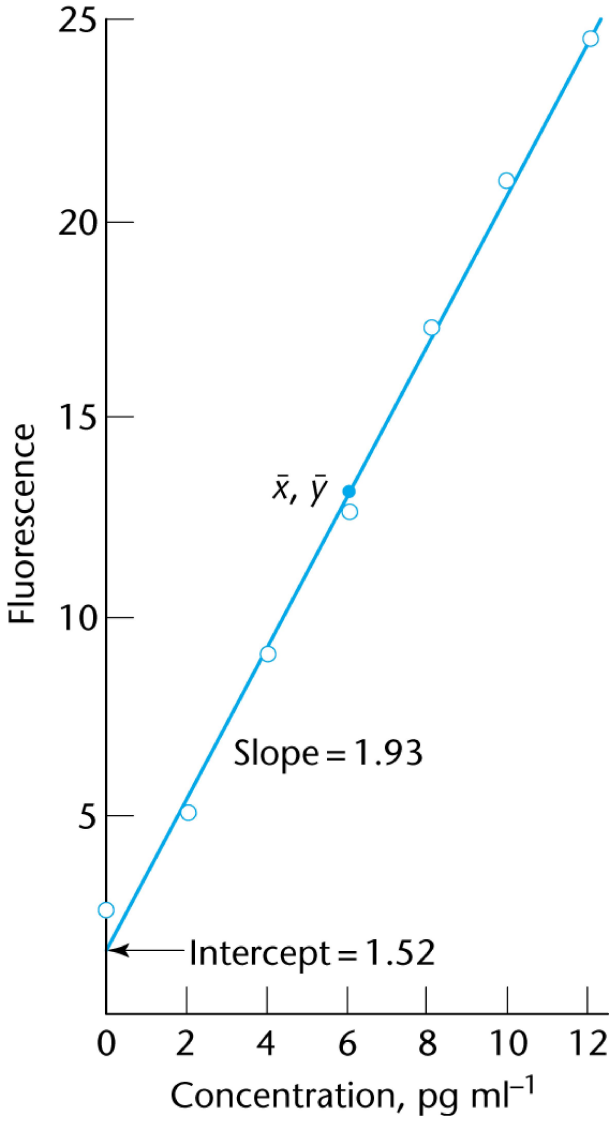
Computations in **R**

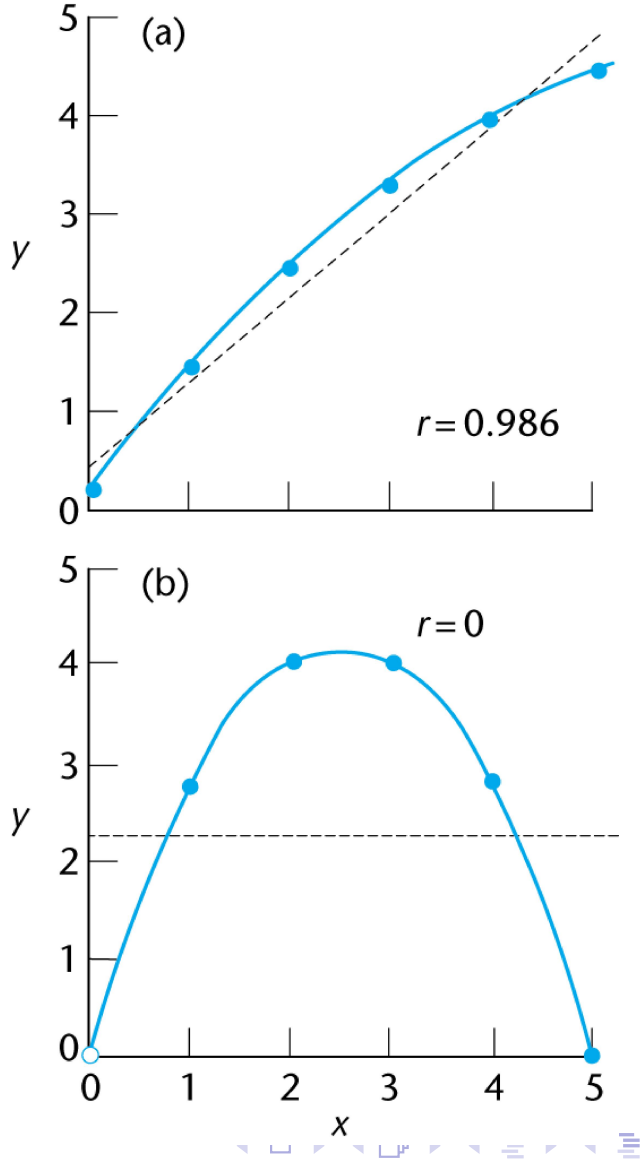
```
Int=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
Conc=c(0,2,4,6,8,10,12)
cor(Int,Conc)
[1] 0.9988796
```

Regression plot

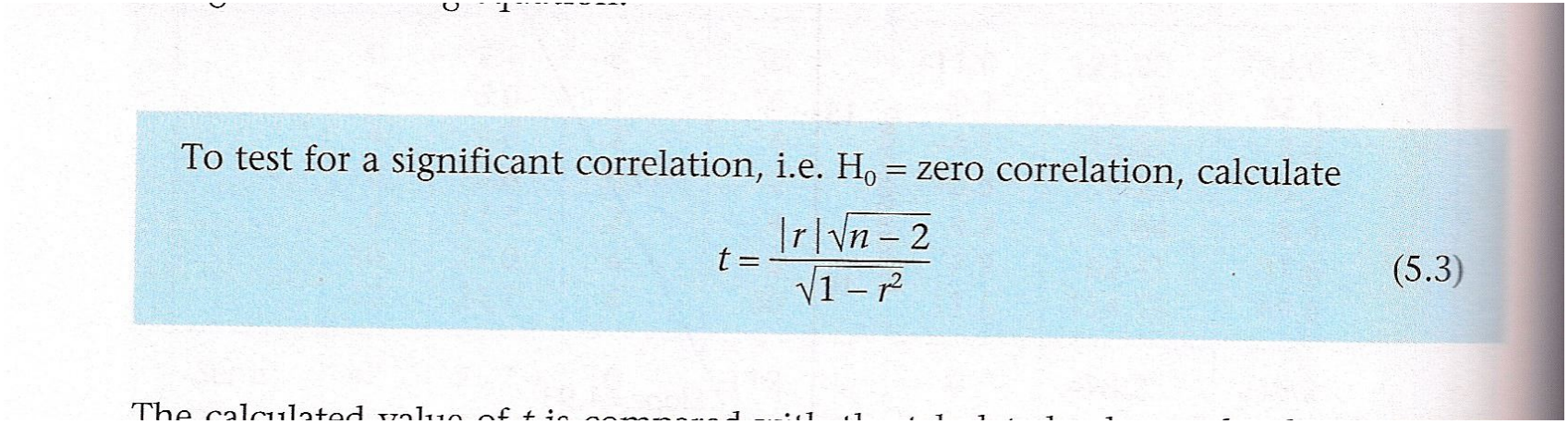


Regression plot





Test for a significance of **R**



To test for a significant correlation, i.e. H_0 = zero correlation, calculate

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (5.3)$$

The calculated value of t is compared with

```
r=cor(Int,Conc)
n=length(Int)
t=abs(r)*sqrt(n-2)/sqrt(1-r^2)
2*(1-pt(t,n-2))
  8.066023e-08
qt(0.975,n-2)
 2.570582
```

More advanced analysis

```
 summary(lm(Int~Conc))

Call:
lm(formula = Int ~ Conc)

Residuals:
       1        2        3        4        5        6        7
 0.58214 -0.37857 -0.23929 -0.50000  0.33929  0.17857  0.01786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5179     0.2949   5.146  0.00363 **
Conc          1.9304     0.0409  47.197 8.07e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-squared: 0.9978,     Adjusted R-squared: 0.9973
F-statistic:  2228 on 1 and 5 DF,  p-value: 8.066e-08
```

Fitting the regression line

It can be shown that the least squares straight line is given by:

$$\text{Slope of least squares line: } b = \frac{\sum_i \{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_i (x_i - \bar{x})^2}$$

$$\text{Intercept of least squares line: } a = \bar{y} - b\bar{x}$$

Example 5.4.1

Calculate the slope and intercept of the regression line for the data given in the previous example (see Section 5.3).

In Section 5.3 we calculated that, for this calibration curve:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 216.2; \qquad \sum_i (x_i - \bar{x})^2 = 112; \qquad \bar{x} = 6; \qquad \bar{y} = 13.1$$

Using equations (5.4) and (5.5) we calculate that

$$b = 216.2/112 = 1.93$$

$$a = 13.1 - (1.93 \times 6) = 13.1 - 11.58 = 1.52$$

The equation for the regression line is thus $y = 1.93x + 1.52$.

Errors in the slope and intercept

$$s_{y/x} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} \qquad (5.6)$$

for the slope (b) and the intercept (a). These are given by:

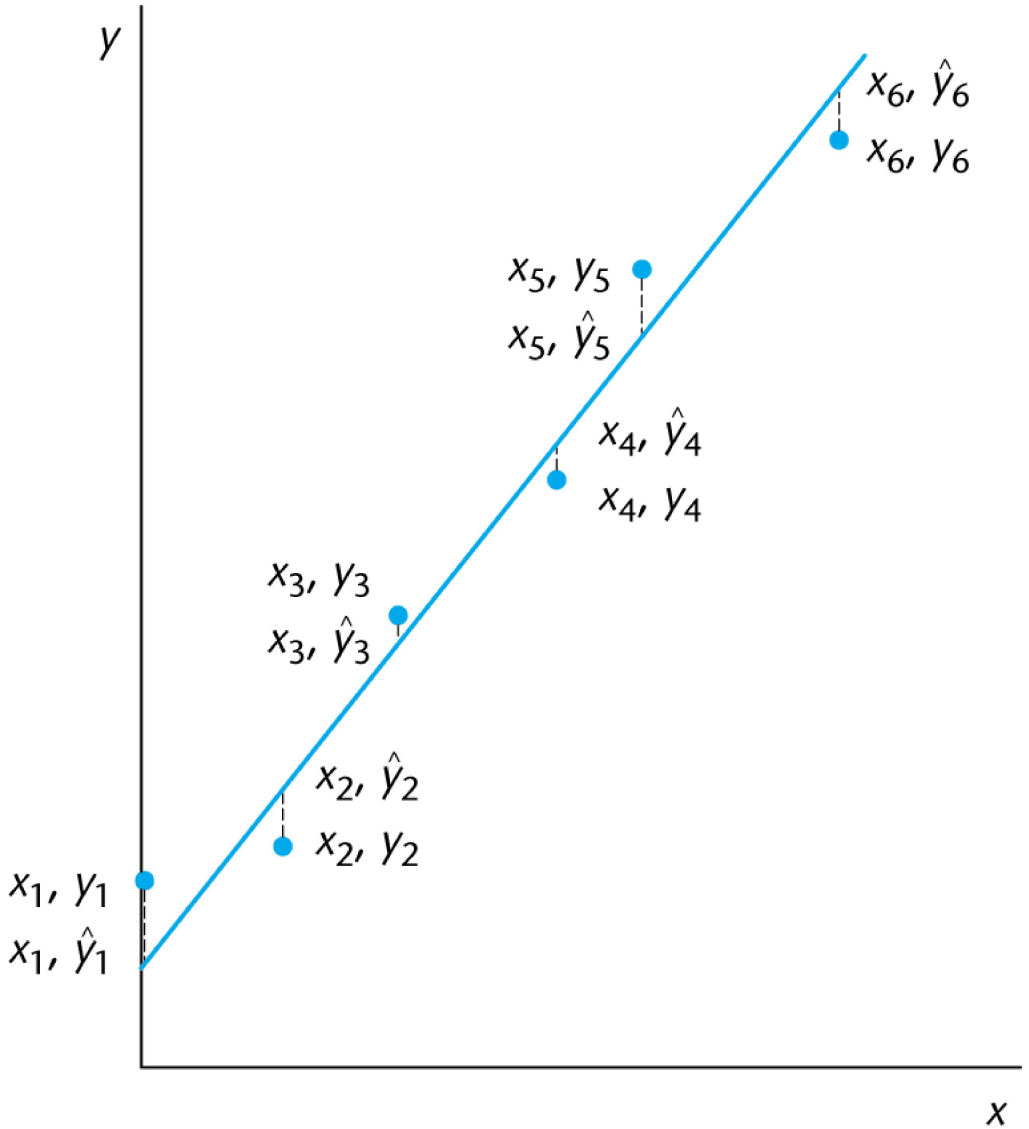
Standard deviation of slope: $$s_b = \frac{s_{y/x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Standard deviation of intercept: $$s_a = s_{y/x}\sqrt{\frac{\sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}}$$

Distribution of the normalized errors

The errors of estimation of the slope and intercept, when normalized by the corresponding standard deviations are distributed as Student t-distribution with $n - 2$ degrees of freedom.

Residuals


y
x
$x_6, \hat{y}_6$
x_6, y_6
x_5, y_5
$x_5, \hat{y}_5$
$x_4, \hat{y}_4$
x_4, y_4
x_3, y_3
$x_3, \hat{y}_3$
$x_2, \hat{y}_2$
x_2, y_2
x_1, y_1
$x_1, \hat{y}_1$

Example

by $a \pm t_{(n-2)} s_a$.

Example 5.5.1

Calculate the standard deviations and confidence limits of the slope and intercept of the regression line calculated in Section 5.4.

This calculation may not be accessible on a simple calculator, but suitable computer software is available. Here we perform the calculation manually, using a tabular layout.

x_i	x_i^2	y_i	$\hat{y}_i$	$\lvert y_i - \hat{y}_i \rvert$	$(y_i - \hat{y}_i)^2$
0	0	2.1	1.52	0.58	0.3364
2	4	5.0	5.38	0.38	0.1444
4	16	9.0	9.24	0.24	0.0576
6	36	12.6	13.10	0.50	0.2500
8	64	17.3	16.96	0.34	0.1156
10	100	21.0	20.82	0.18	0.0324
12	144	24.7	24.68	0.02	0.0004
	$\sum_i x_i^2 = 364$				$\sum_i (y_i - \hat{y}_i)^2 = 0.9368$

From the table and using equation (5.6) we obtain

$$s_{y/x} = \sqrt{0.9368/5} = \sqrt{0.18736} = 0.4329$$

From Section 5.3 we know that $\sum_i (x_i - \bar{x})^2 = 112$, and equation (5.7) can be used to show that

$$s_b = 0.4329/\sqrt{112} = 0.4329/10.58 = 0.0409$$

The t-value for $(n - 2) = 5$ degrees of freedom and the 95% confidence level is 2.57 (Table A.2). The 95% confidence limits for b are thus:

$$b = 1.93 \pm (2.57 \times 0.0409) = 1.93 \pm 0.11$$

Equation (5.8) requires knowledge of $\sum_i x_i^2$, calculated as 364 from the table. We can thus write:

$$s_a = 0.4329\sqrt{\frac{364}{7 \times 112}} = 0.2950$$

so the 95% confidence limits are:

$$a = 1.52 \pm (2.57 \times 0.2950) = 1.52 \pm 0.76$$