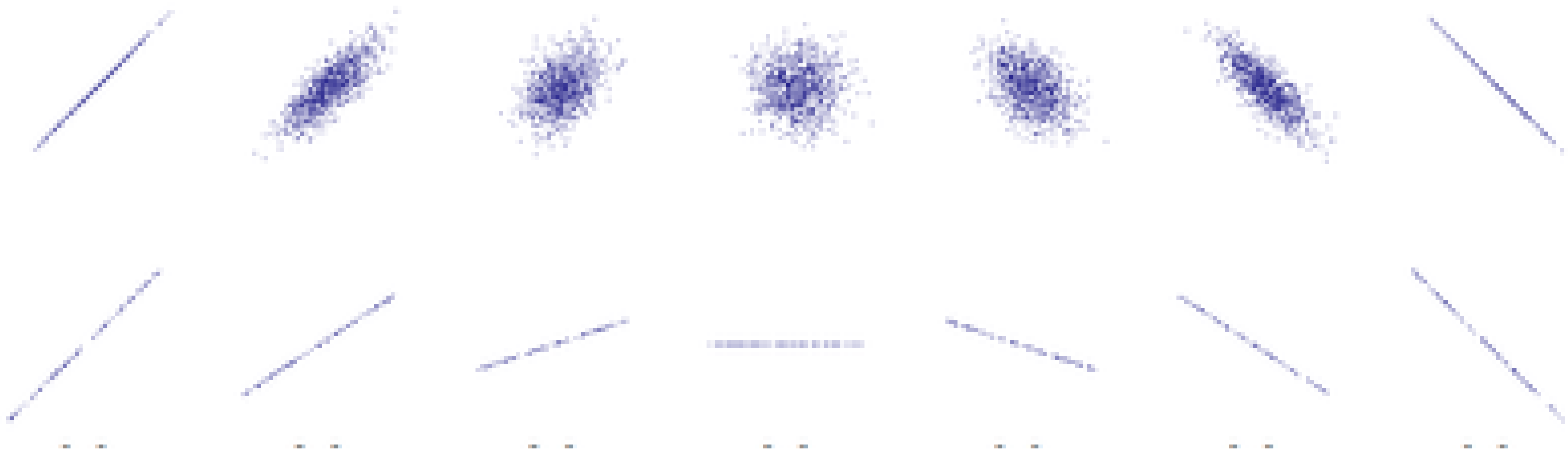
FINDING RELATIONS BETWEEN VARIABLES

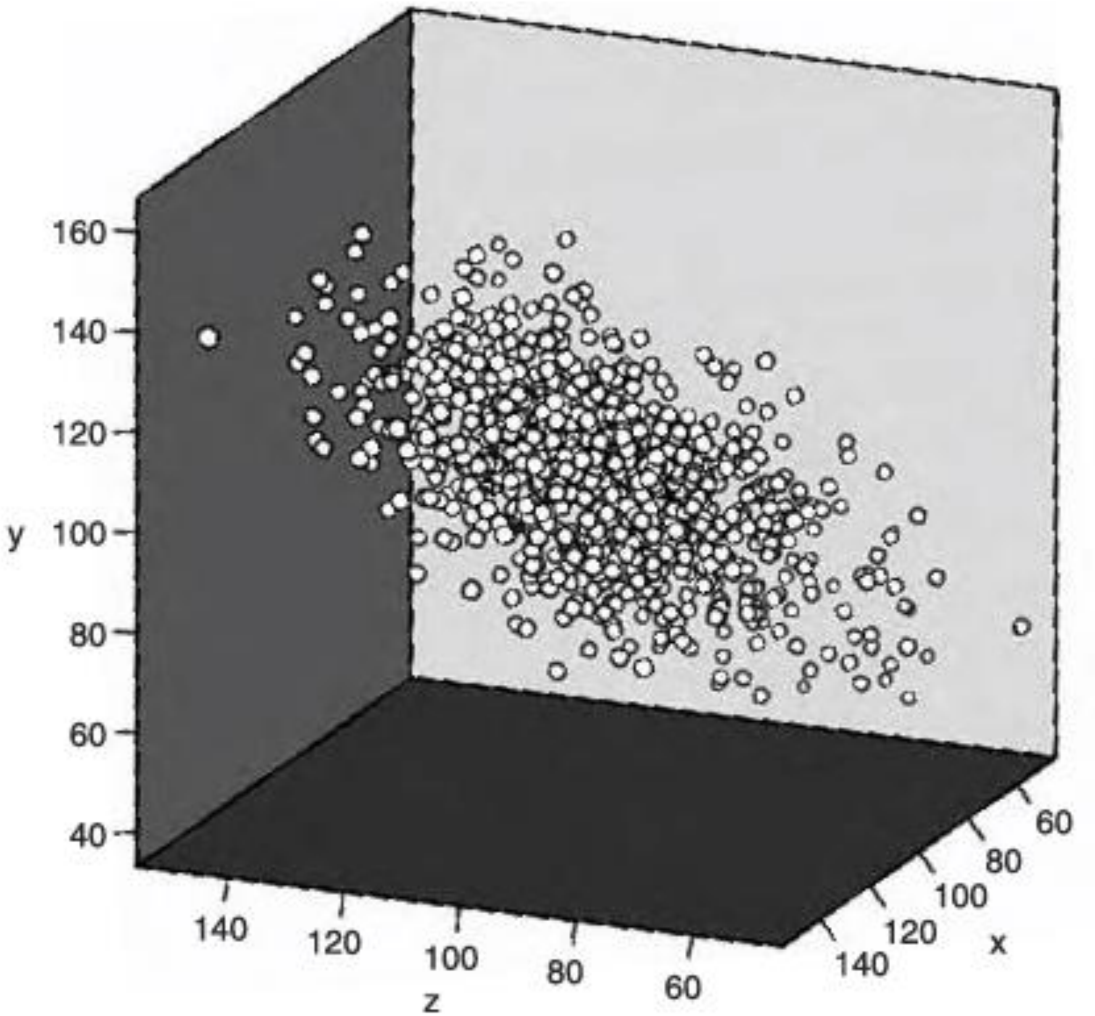
Correlation

Relation between coupled variables

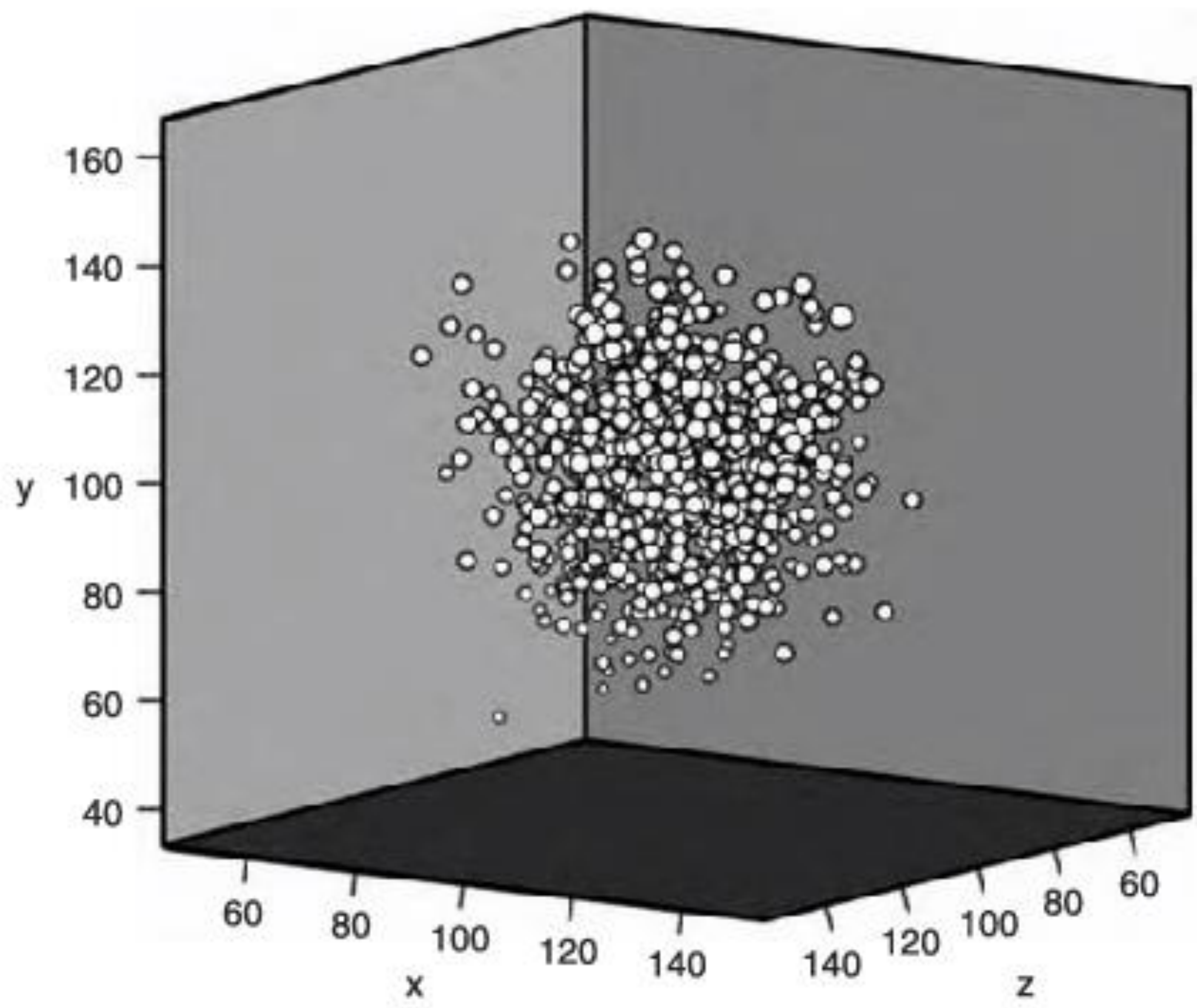


What couples of variables are in relation?

Correlated variables



Uncorrelated variables



Variance and Moments of a Random Variable

- **Definition**
  - The covariance of two random variable X and Y is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

- **Theorem**
  - For any two random variables X and Y.

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y).$$

Independent variables→ COV=0

$$Cov(X,Y) = E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] =$$
$$= E[XY] - E[X]E[Y]$$

$E[XY] =$ For discrete variables (for continuous, integral instead of sum)

$= \sum_{i,j} x_i y_j P(x_i y_j) =$ For Independent variables

$$= \sum_{i,j} x_i y_j P(x_i) P(y_j) = \sum_i x_i P(x_i) \sum_j y_j P(y_j) = E[X]E[Y]$$

X,Y independent → COV (X,Y) =0

The viceversa is not always true

Covariance and Pearson's Correlation index

	Variable 1	Variable 2
Item1	x_{11}	x_{21}
Item 2	x_{12}	x_{22}
Item i	x_{1i}	x_{2i}
Item m	x_{1m}	x_{2m}
Mean	M_{1-}	M_{2-}

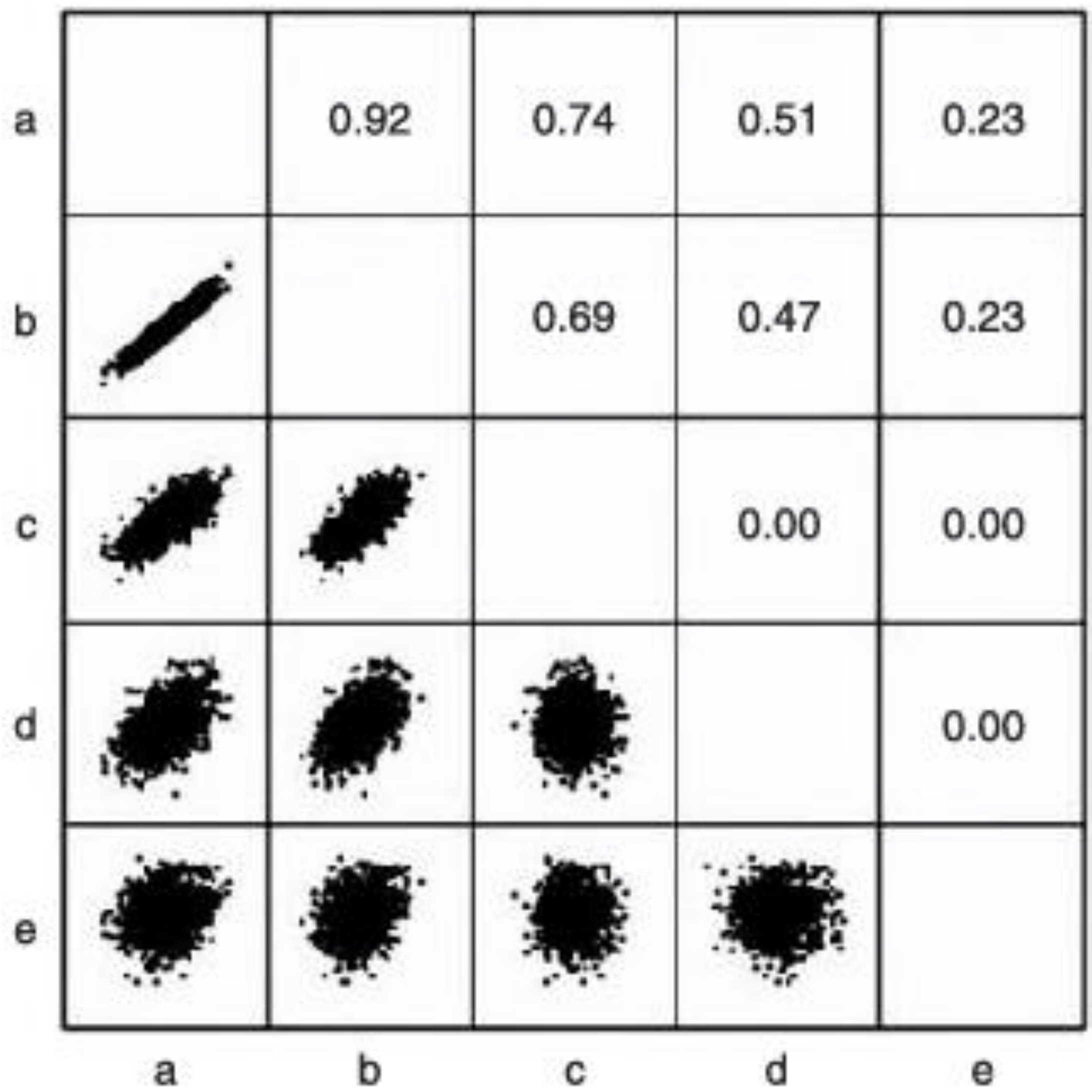
$$M_{1-} = \frac{1}{n}\sum_{i=1}^{m} x_{1i}$$

$$M_{2-} = \frac{1}{n}\sum_{i=1}^{m} x_{2i}$$

$$\mathrm{cov}(x_{1-}, x_{2-}) = \frac{1}{n-1}\sum_{i=1}^{m}\left(x_{1i} - M_{1-}\right)\left(x_{2i} - M_{2-}\right)$$

$$corr(x_{1-}, x_{2-}) = \frac{1}{n-1}\sum_{i=1}^{m}\frac{\left(x_{1i} - M_{1-}\right)\left(x_{2i} - M_{2-}\right)}{\sigma_{1-}\sigma_{2-}}$$

Correlation



$$corr(x_1, x_2) \in [-1,1]$$

When is a correlation significant?

Given a correlation index:

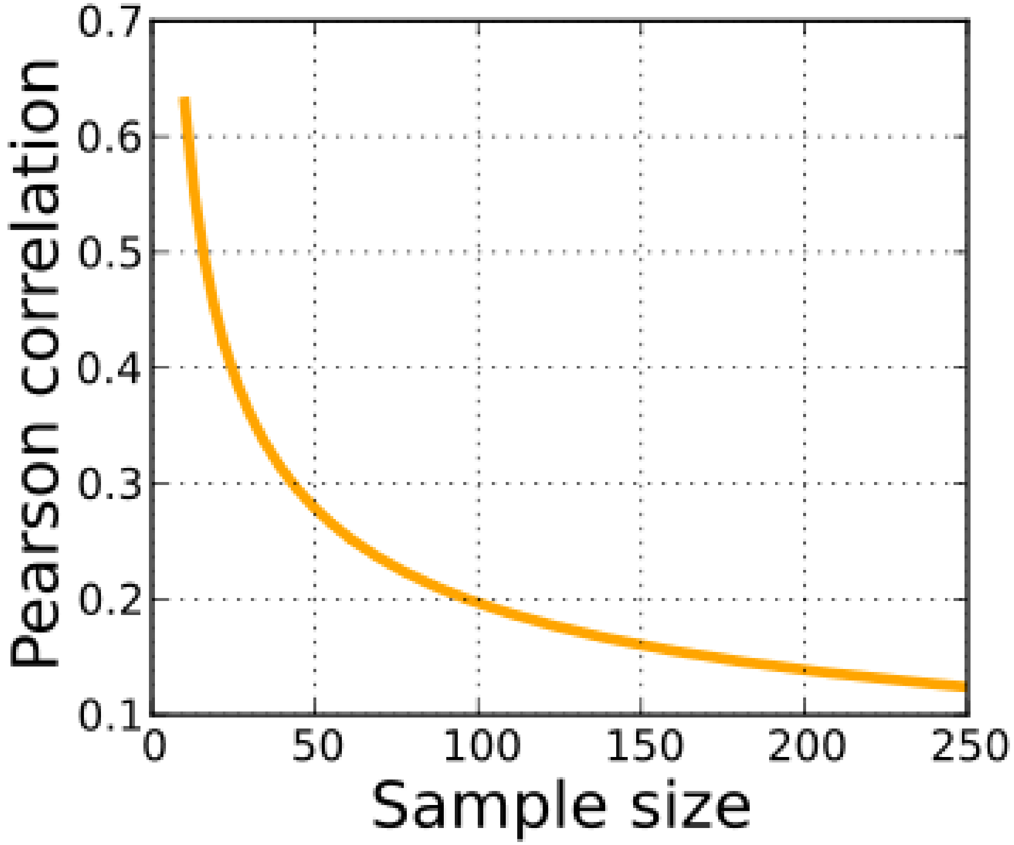
$$r(x_{1-}, x_{2-}) = \frac{1}{n-1}\sum_{i=1}^{m}\frac{\left(x_{1i} - M_{1-}\right)\left(x_{2i} - M_{2-}\right)}{\sigma_{1-}\sigma_{2-}}$$

A test variable can be computed under the null hypothesis that r=0

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

t is distributed as Student's t test with n-2 degrees of freedom

It assumes normality of x



Graph showing the minimum value of Pearson's correlation coefficient that is significantly different from zero at the 0.05 level, for a given sample size.

Example: discovery of a misconduct

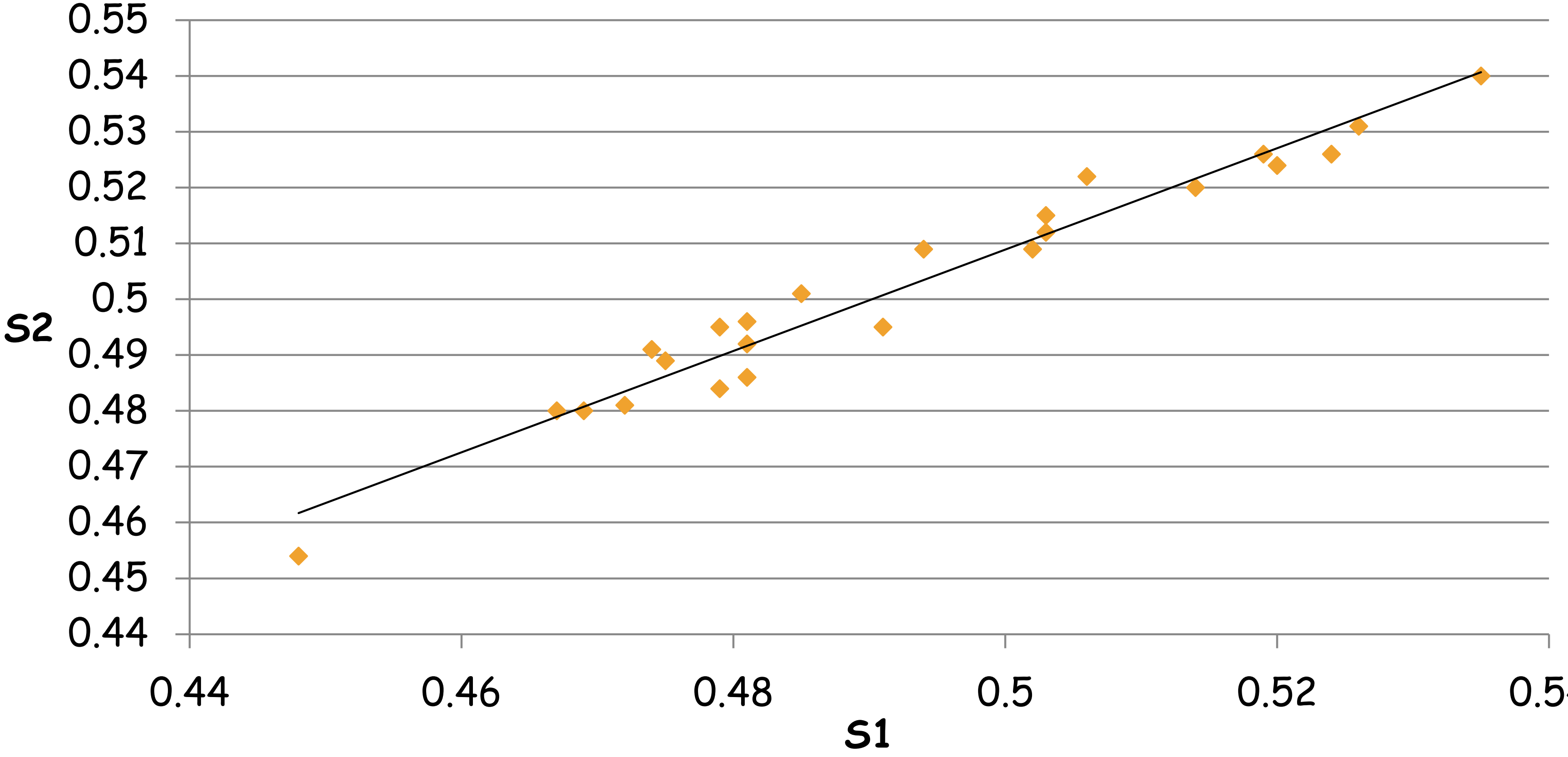
- Repeatability test: 2 different experimentalist were asked to take the same solution and to perform 24 independent ELISA assays on a 6x4 plate.

- They submitted to the assessor the following results out of the spectrophotometer, ordered following the well

	S1	S2
P1	0,481	0,496
P2	0,485	0,501
P3	0,479	0,495
P4	0,506	0,522
P5	0,467	0,48
P6	0,474	0,491
P7	0,469	0,48
P8	0,475	0,489
P9	0,514	0,52
P10	0,52	0,524
P11	0,526	0,531
P12	0,494	0,509
P13	0,535	0,54
P14	0,524	0,526
P15	0,481	0,492
P16	0,502	0,509
P17	0,479	0,484
P18	0,491	0,495
P19	0,503	0,515
P20	0,472	0,481
P21	0,481	0,486
P22	0,503	0,512
P23	0,448	0,454
P24	0,519	0,526

- The assessor suspected that the experimenter submitted two reads of the same plate
- How to prove it?

Example: discovery of a misconduct



R=0.978

Example: discovery of a misconduct

- R=0.978 n=24 → t=22.05

t(p,df)

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251

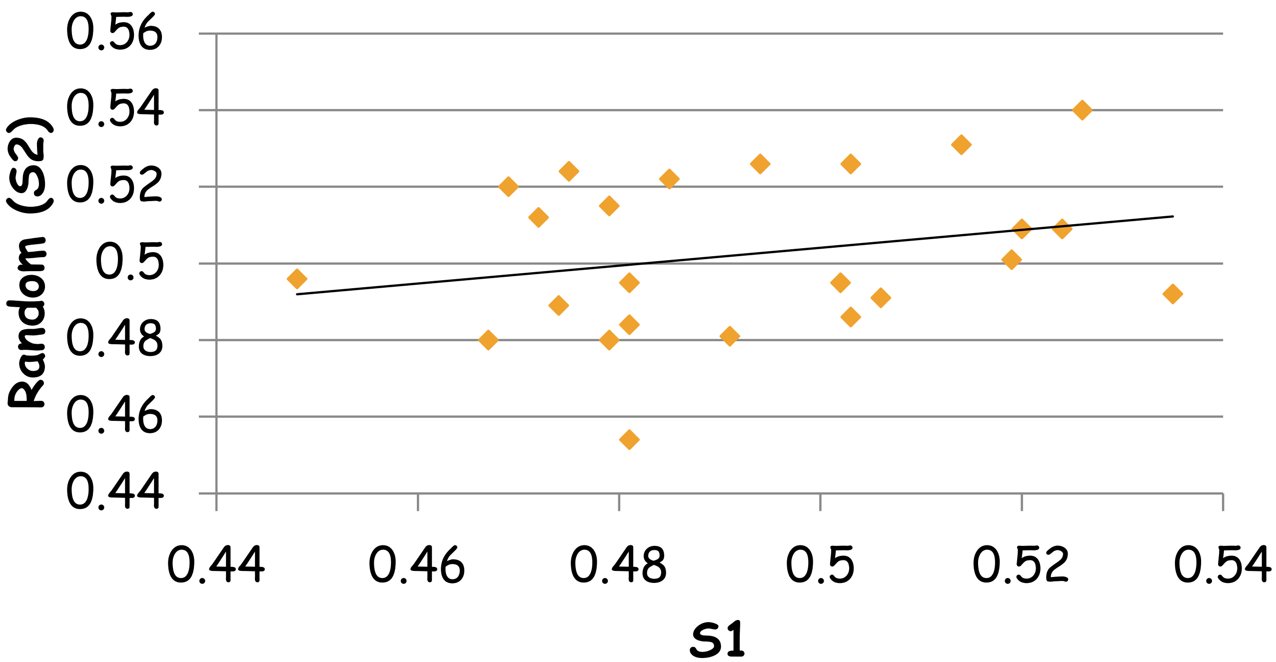
Objection: the test is valid only when data are normally distributed and we cannot prove that.

Any other idea?

	S1	S2
P1	0,481	0,496
P2	0,485	0,501
P3	0,479	0,495
P4	0,506	0,522
P5	0,467	0,48
P6	0,474	0,491
P7	0,469	0,48
P8	0,475	0,489
P9	0,514	0,52
P10	0,52	0,524
P11	0,526	0,531
P12	0,494	0,509
P13	0,535	0,54
P14	0,524	0,526
P15	0,481	0,492
P16	0,502	0,509
P17	0,479	0,484
P18	0,491	0,495
P19	0,503	0,515
P20	0,472	0,481
P21	0,481	0,486
P22	0,503	0,512
P23	0,448	0,454
P24	0,519	0,526

Use the data themselves to generate random experiments

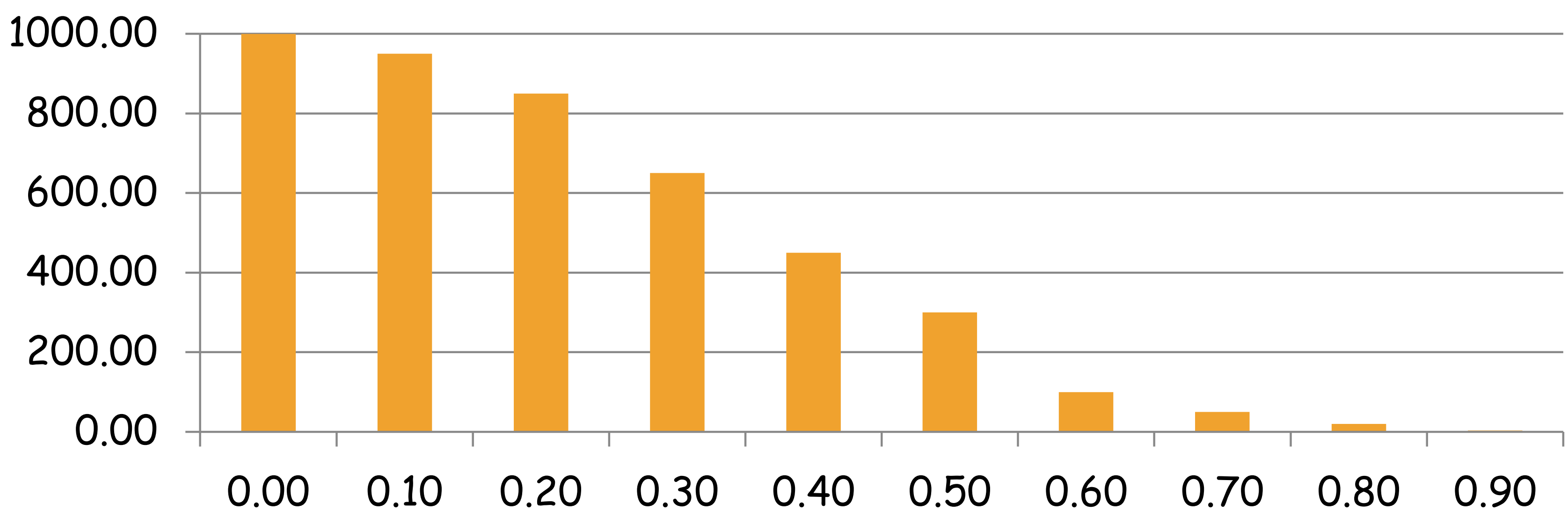
	S1	Random(S2)
P1	0,481	0,495
P2	0,485	0,522
P3	0,479	0,48
P4	0,506	0,491
P5	0,467	0,48
P6	0,474	0,489
P7	0,469	0,52
P8	0,475	0,524
P9	0,514	0,531
P10	0,52	0,509
P11	0,526	0,54
P12	0,494	0,526
P13	0,535	0,492
P14	0,524	0,509
P15	0,481	0,484
P16	0,502	0,495
P17	0,479	0,515
P18	0,491	0,481
P19	0,503	0,486
P20	0,472	0,512
P21	0,481	0,454
P22	0,503	0,526
P23	0,448	0,496
P24	0,519	0,501



R=0.25

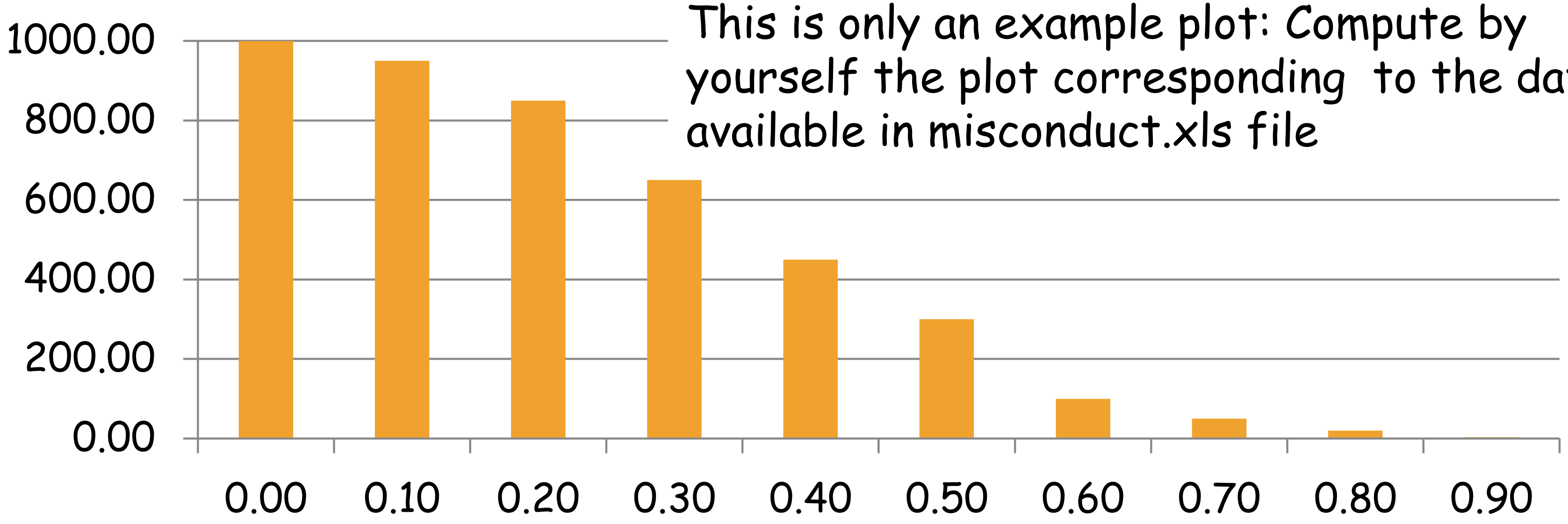
Building a distribution by random resampling

- Iterate the process of shuffling and computation of r many times (say 1000)
- Compute a cumulative histogram counting the resamplings scoring with correlation≥r



Building a distribution by random resampling

This is only an example plot: Compute by yourself the plot corresponding to the data available in misconduct.xls file



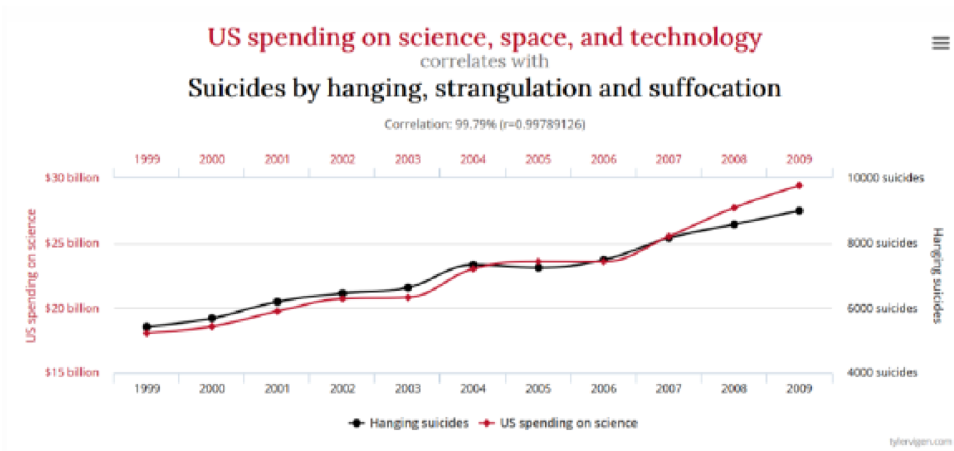
- The plot gives the probability (per thousand) of obtaining a given correlation with random pairings of the original data→ P-value independent on the assumptions on the data distribution

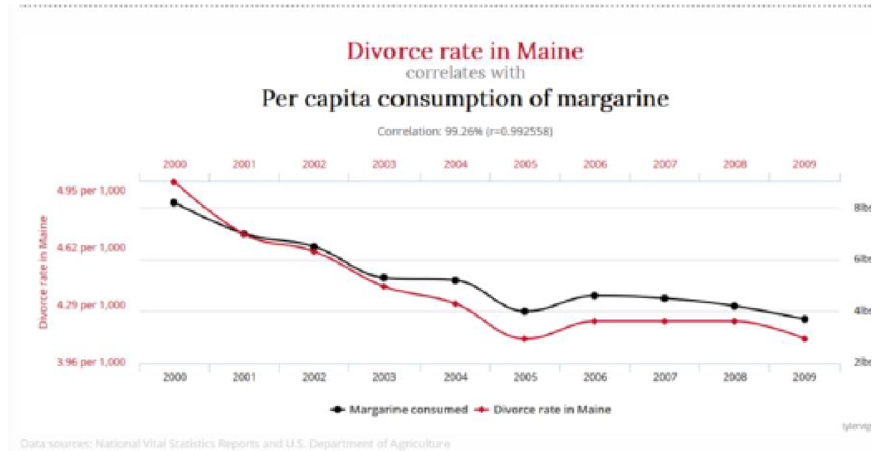
Bootstrapping

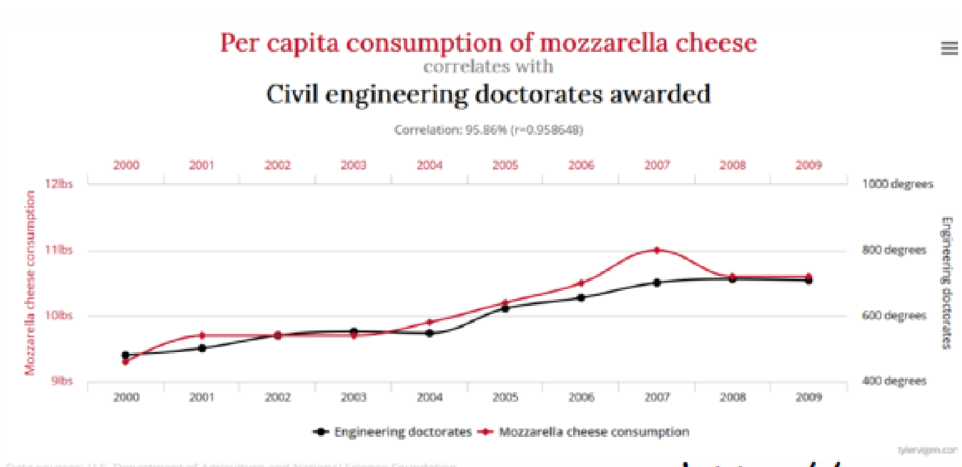
- The term is often attributed to Rudolf Erich Raspe's story *The Surprising Adventures of Baron Munchausen*, where the main character pulls himself out of a swamp by his hair (specifically, his pigtail), but the Baron does not, in fact, pull himself out by his bootstraps

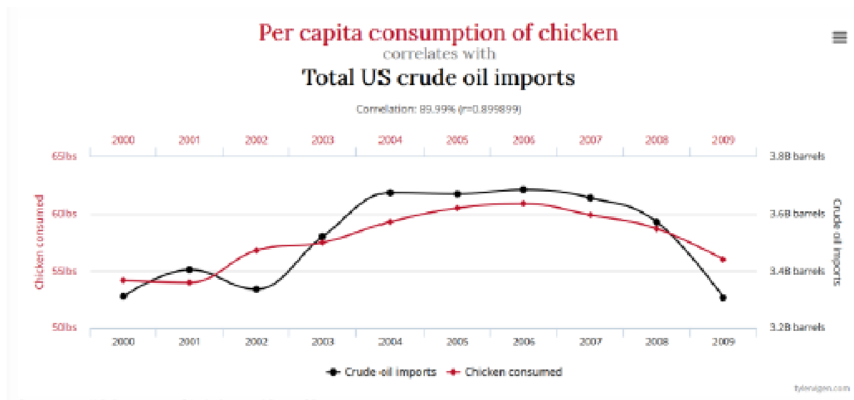
Correlation is not causation

- Spurious correlations









http://www.tylervigen.com/spurious-correlations

Correlation means a co-relation is observed, which does not imply a casual relation.

- If X and Y are strongly correlated, this may have many reasons.
- Besides spurious, it may be that X and Y are the result of an unobserved process Z.

In both cases X is correlated to Y, BUT there is not direct causal relation

IF YOU PICK ALL THE CASE WHEN Z IS CONSTANT THERE WILL BE NO CORRELATION BETWEEN X AND Y

Partial correlation coefficient should be computed (it is possible only when Z is known)

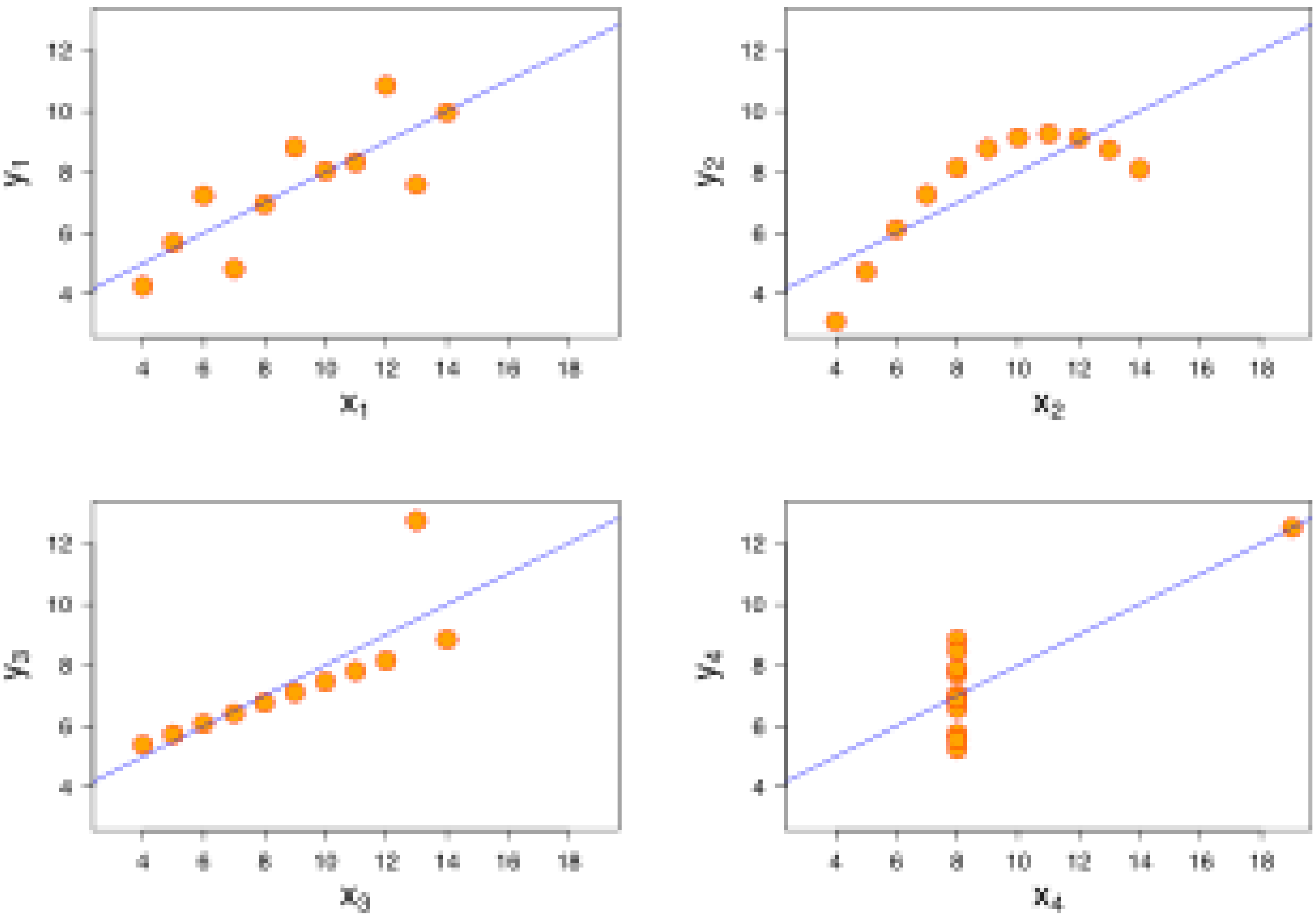
Precision matrix (we'll see better next year)

- Given a set of sample describe by variables

  X_1 X_2 X_3 X_4 … X_N
- Compute the Covariance Matrix
- Compute the Precision Matrix (K) as the **inverse of the covariance matrix**
- The partial correlation indexes between pairs of variables X_i X_j, with $i \neq j$ is

$$\tilde{\rho}\left(X_i, X_j\right) = \frac{-K_{ij}}{\sqrt{K_{ii} K_{jj}}}$$

Correlation index assumes linear dependence

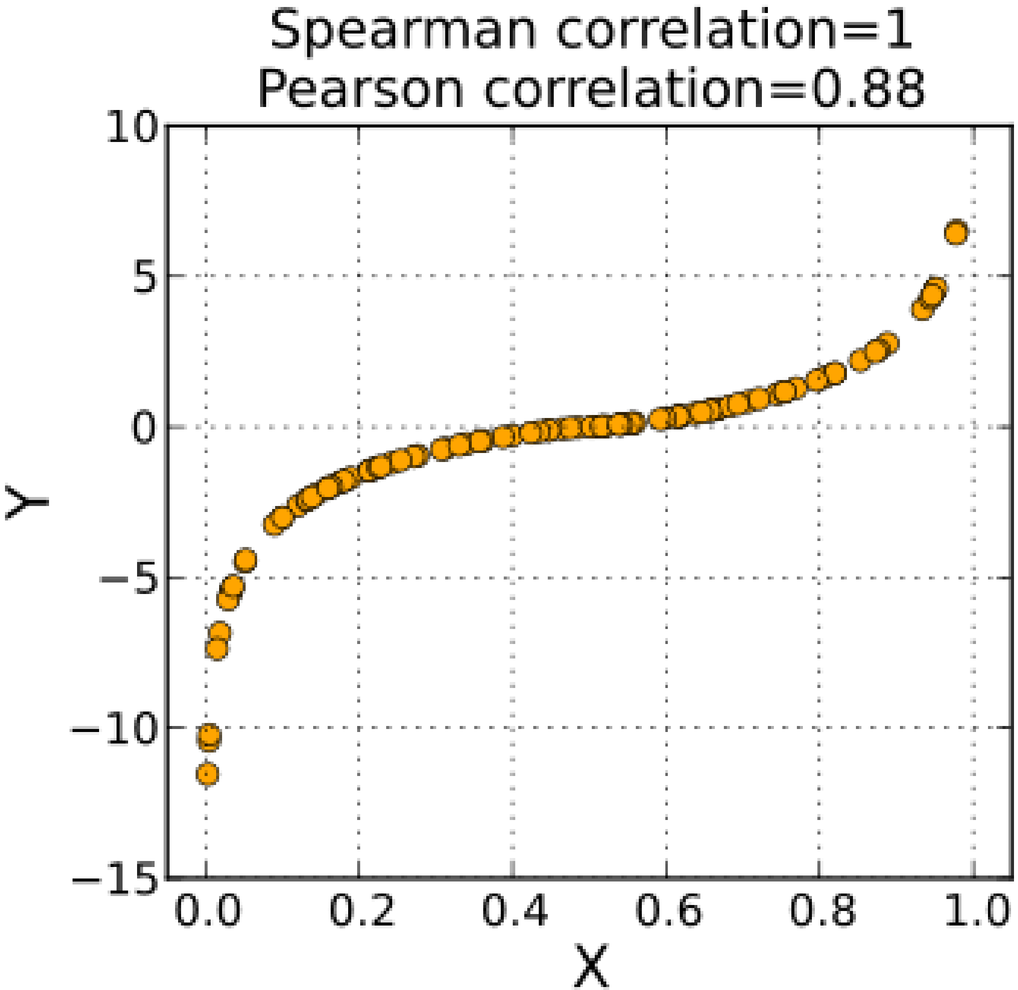


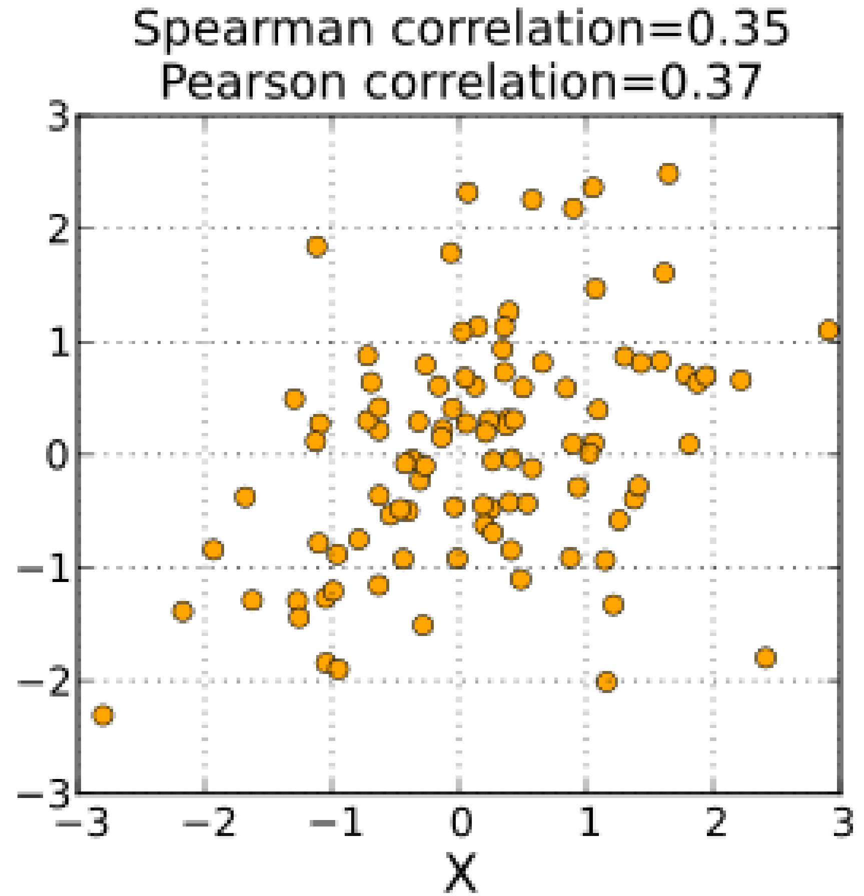
R=0.816

Non parametric correlation: Spearman

- Given a set of paired (x_i, y_i) sort separately the two variables, obtaining the ranks.

- The Sperman's correlation is the Pearson's correlation of the ranked variables: (Rx_i, Ry_i),


Spearman correlation=1
Pearson correlation=0.88
Y
X


Spearman correlation=0.35
Pearson correlation=0.37
X

- Under the null hypothesis (r=0)

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Is distributed as a Student's t test with n-2 degrees of freedom

Categorical data: Matthews correlation index

	Secreted	Non Secreted	Total
With Signal peptide	a	b	a + b
Without Signal Peptide	c	d	c + d
total	a + c	b + d	n

$$MCC = \frac{ad\text{-}bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$

Indexes for general dependence

- Mutual information

$$I[X,Y] = E_{x,y}\left[\ln \frac{p(x,y)}{p(x)p(y)}\right]$$

Discrete $$I[X,Y] = \sum_{x_i}\sum_{y_j} p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Continous $$I[X,Y] = \int\int p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} dxdy$$

Indexes for general dependence

- Mutual information

$$I[X,Y]=E_{x,y}\left[\ln\frac{p(x,y)}{p(x)p(y)}\right]=E_{x,y}\left[\ln\frac{p(x)p(y\mid x)}{p(x)p(y)}\right]=$$

$$=E_{x,y}[\ln p(y\mid x)]-E_{x,y}[\ln p(y)]=$$

$$=\int p(x)\left(\int p(y\mid x)\ln p(y\mid x)dy\right)dx-\int p(y)\ln p(y)\left(\int p(x\mid y)dx\right)dy=$$

$$=\int p(x)\left(\int p(y\mid x)\ln p(y\mid x)dy\right)dx-\int p(y)\ln p(y)=$$

$$=-H(Y\mid X)+H(Y)$$

$H(X)=E[\ln x]$ Information (Shannon's) Entropy

In practice: find the optimal discretization: Maximum Information Coefficient

MIC in a pic

Given $D \subset \mathbb{R}^2$ and integers x and y,

$$I^*(D, x, y) = \max I(D|_G)$$

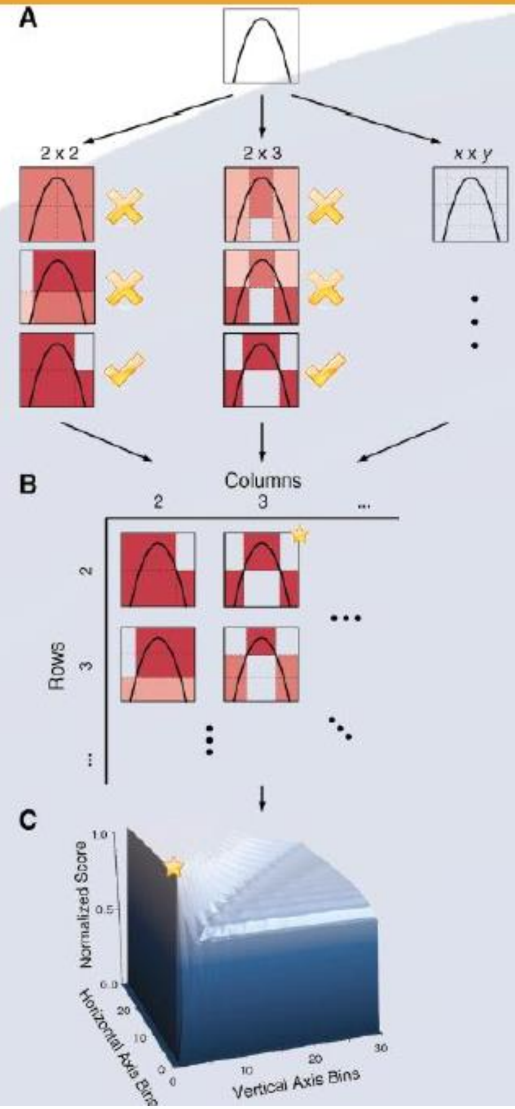
with G over all grids of x cols, y rows.

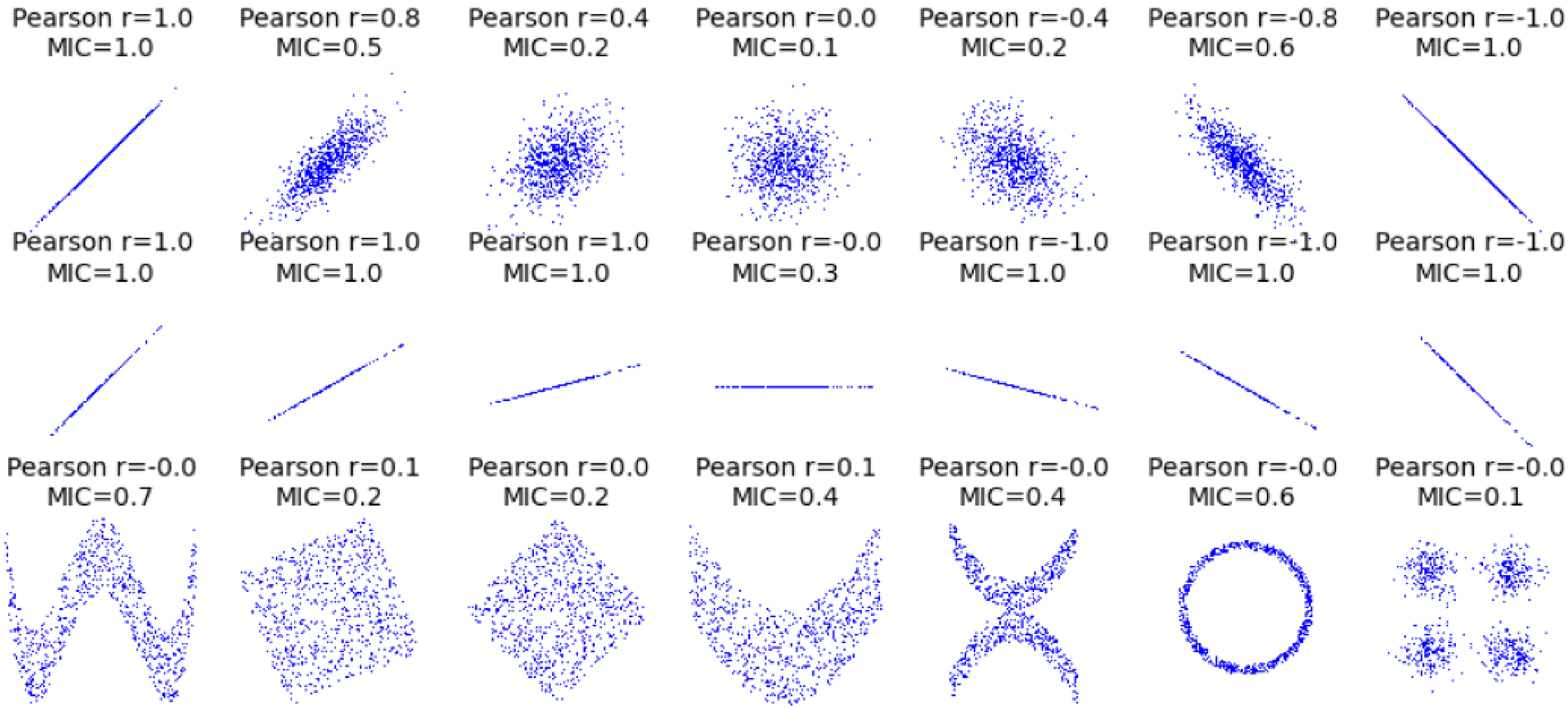
Normalise this score by independence

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min(x, y)}$$

And return the maximum

$$MIC(D) = \max_{xy<B(n)} \{M(D)_{x,y}\}$$

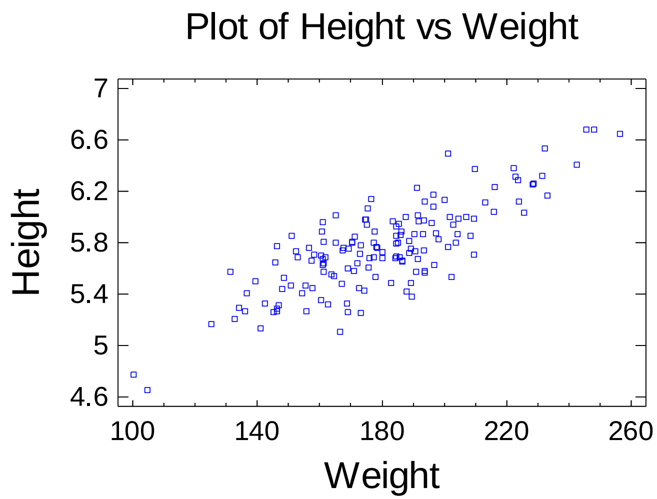



Pearson r=1.0 MIC=1.0
Pearson r=0.8 MIC=0.5
Pearson r=0.4 MIC=0.2
Pearson r=0.0 MIC=0.1
Pearson r=-0.4 MIC=0.2
Pearson r=-0.8 MIC=0.6
Pearson r=-1.0 MIC=1.0
Pearson r=1.0 MIC=1.0
Pearson r=1.0 MIC=1.0
Pearson r=1.0 MIC=1.0
Pearson r=-0.0 MIC=0.3
Pearson r=-1.0 MIC=1.0
Pearson r=-1.0 MIC=1.0
Pearson r=-1.0 MIC=1.0
Pearson r=-0.0 MIC=0.7
Pearson r=0.1 MIC=0.2
Pearson r=0.0 MIC=0.2
Pearson r=0.1 MIC=0.4
Pearson r=-0.0 MIC=0.4
Pearson r=-0.0 MIC=0.6
Pearson r=-0.0 MIC=0.1

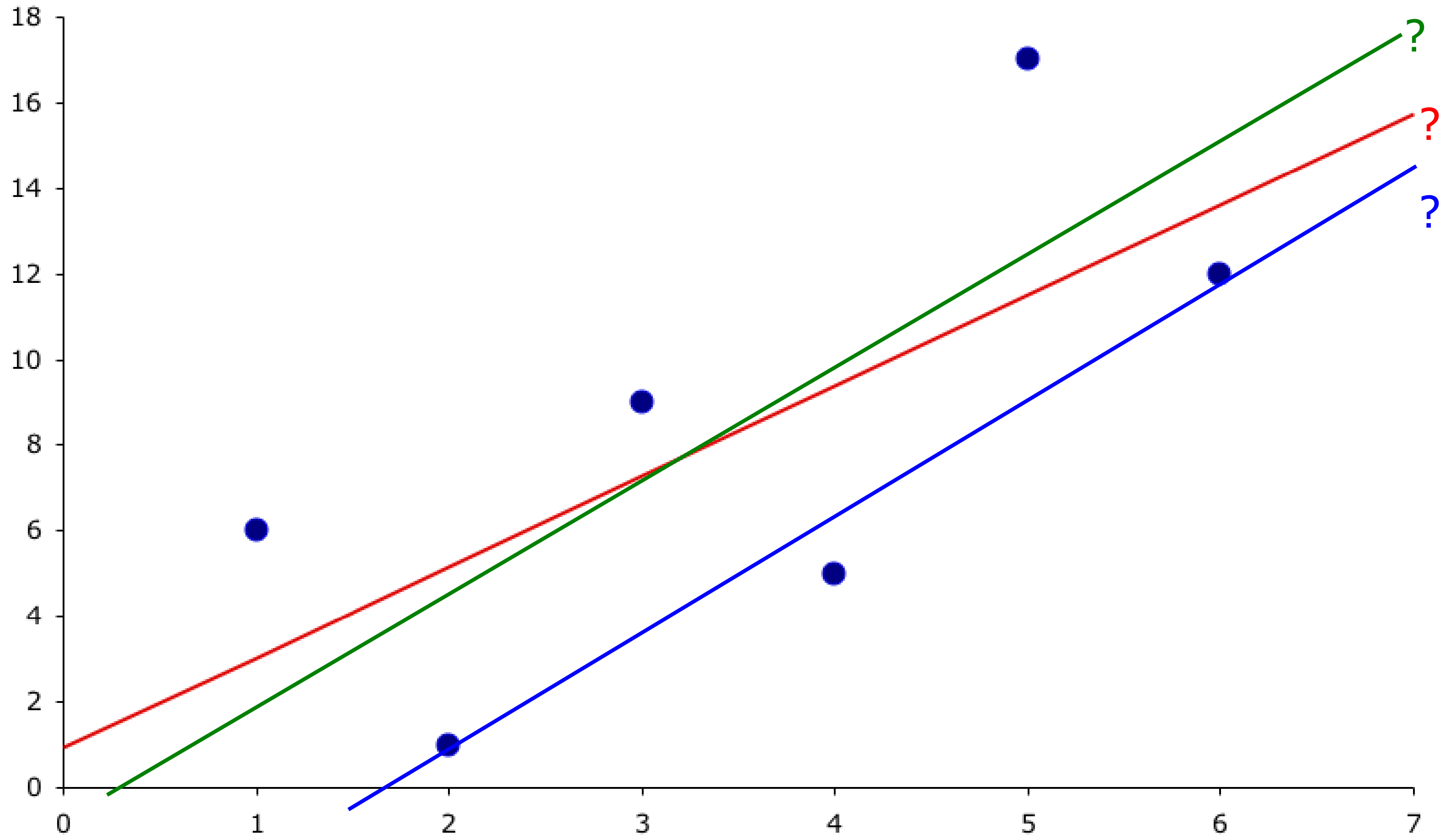
REGRESSION

Regression

- **Regression analysis:** any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.



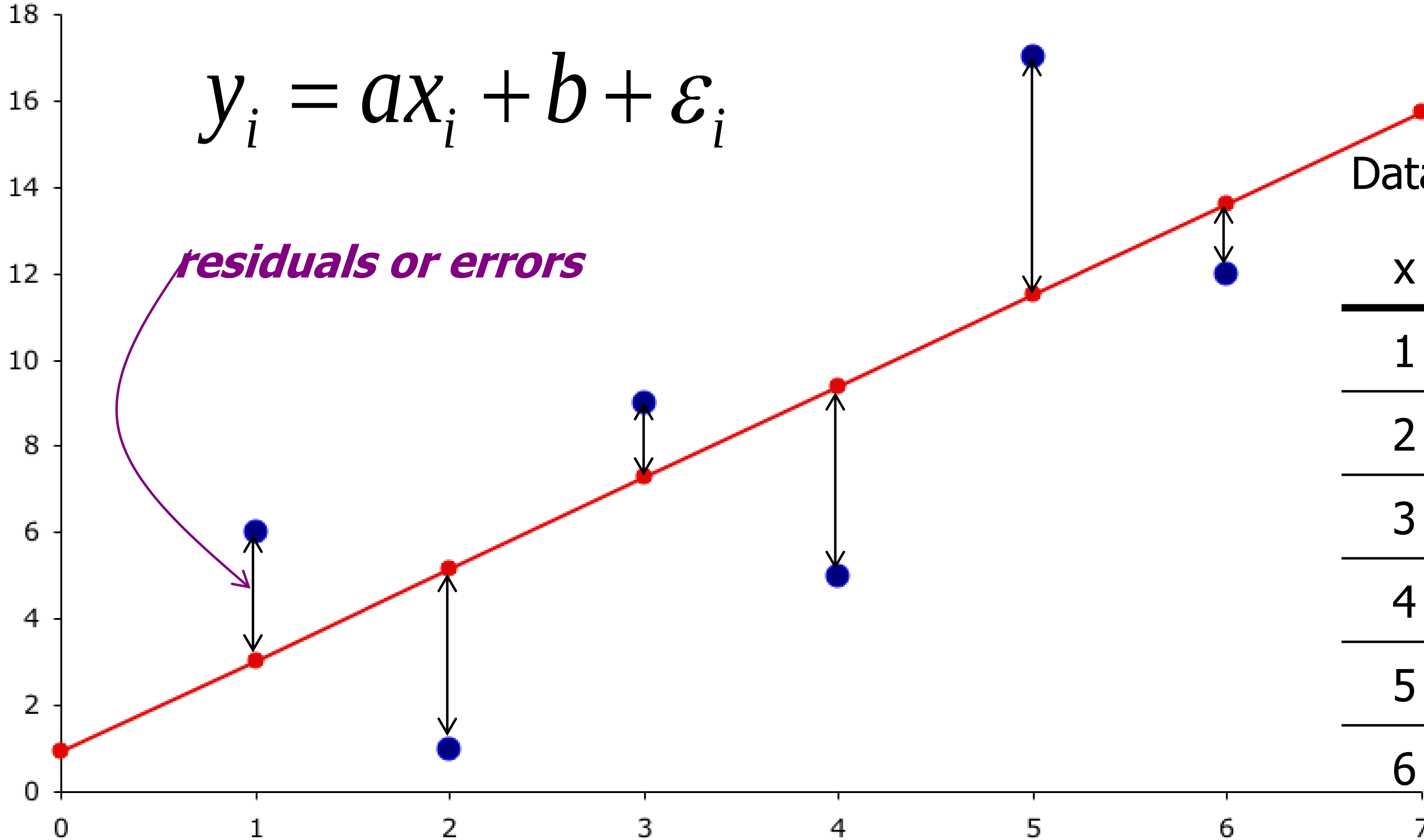
Linear regression: Which is the line that best fits the data?



Linear regression

$$y = ax + b$$ Interpolating line

$$y_i = ax_i + b + \varepsilon_i$$



Data Points:

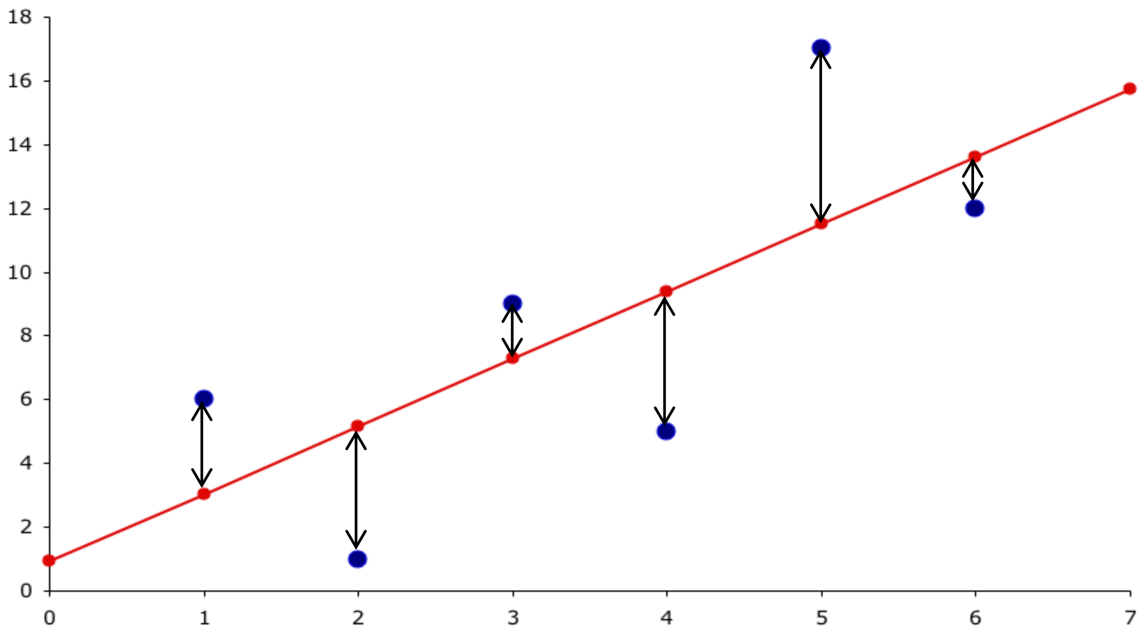
x	y
1	6
2	1
3	9
4	5
5	17
6	12

Least squares line

Choose the line (a,b) that **minimize**

$$y = ax + b$$

$$E = \sum_{i=1}^{m} \varepsilon_i^{\,2} = \sum_{i=1}^{m} \left[y_i - \left(a x_i + b \right) \right]^2$$



Minimizing $E=\sum_{i=1}^{m}\varepsilon_i^2=\sum_{i=1}^{m}\left[y_i-(ax_i+b)\right]^2$

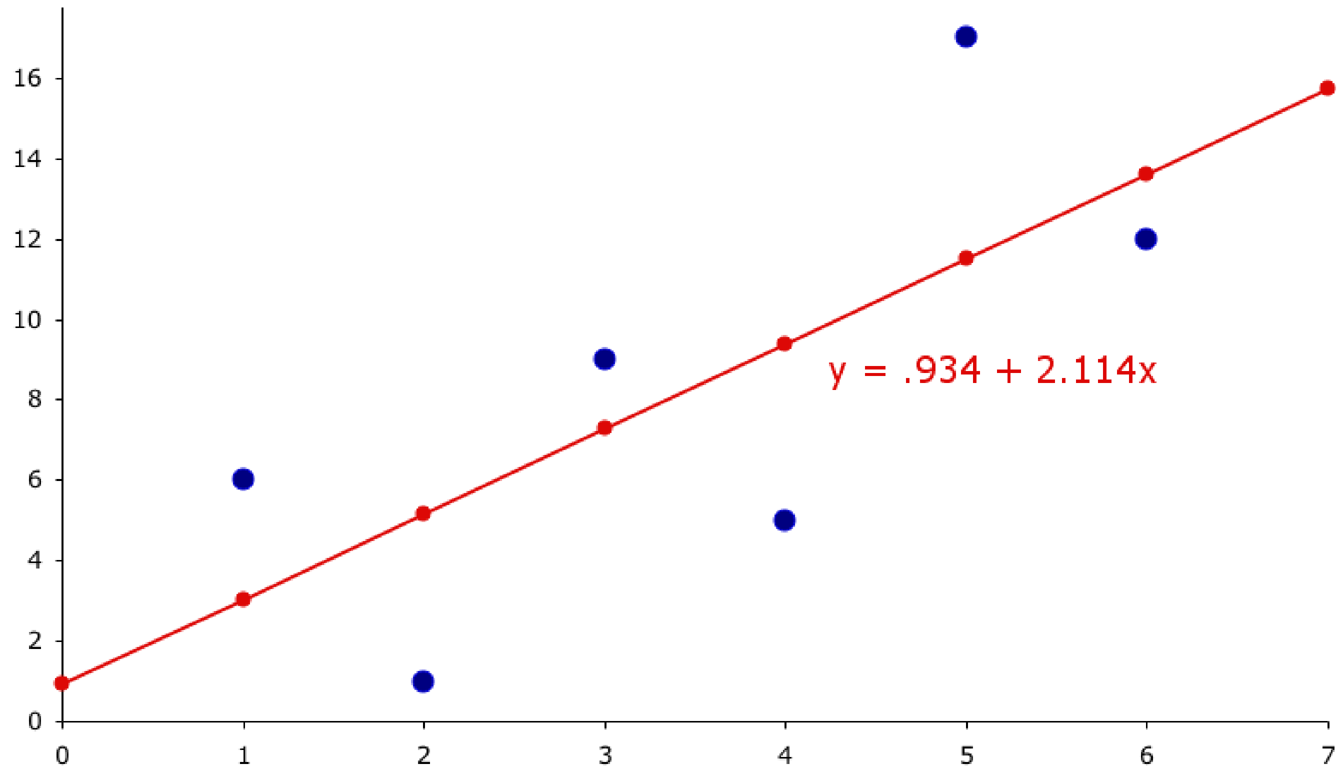
$$\frac{\partial E}{\partial b}=0\Rightarrow 2\sum_{i=1}^{m}\left[y_i-(ax_i+b)\right]=0\Rightarrow b=\frac{1}{m}\sum_{i=1}^{m}y_i-a\frac{1}{m}\sum_{i=1}^{m}x_i\Rightarrow b=\bar{y}-a\bar{x}$$

$$\frac{\partial E}{\partial a}=0\Rightarrow 2\sum_{i=1}^{m}\left[y_i-(ax_i+b)\right]\cdot x_i=0\Rightarrow \sum_{i=1}^{m}x_iy_i-a\sum_{i=1}^{m}x_i^2-\bar{y}\sum_{i=1}^{m}x_i+a\bar{x}\sum_{i=1}^{m}x_i=0\Rightarrow$$

$$\Rightarrow a=\frac{\sum_{i=1}^{m}x_iy_i-\bar{y}\sum_{i=1}^{m}x_i}{\sum_{i=1}^{m}x_i^2-\bar{x}\sum_{i=1}^{m}x_i}=\frac{\sum_{i=1}^{m}x_iy_i-m\bar{y}\bar{x}}{\sum_{i=1}^{m}x_i^2-m\bar{x}\bar{x}}=\frac{\frac{1}{m}\sum_{i=1}^{m}x_iy_i-\bar{y}\bar{x}}{\frac{1}{m}\sum_{i=1}^{m}x_i^2-\bar{x}\bar{x}}\Rightarrow a=\frac{\mathrm{cov}(x,y)}{\sigma_x^2}$$

Data Points:

x	y
1	6
2	1
3	9
4	5
5	17
6	12



Polynomial interpolation

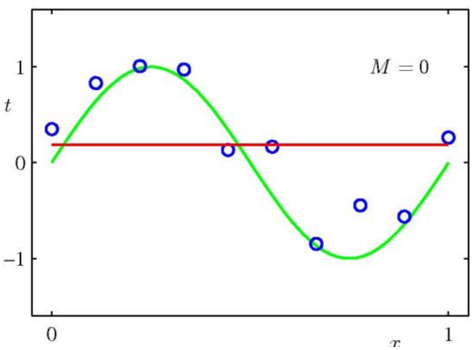
- The same technique can be applied by imposing a polynomial regression model

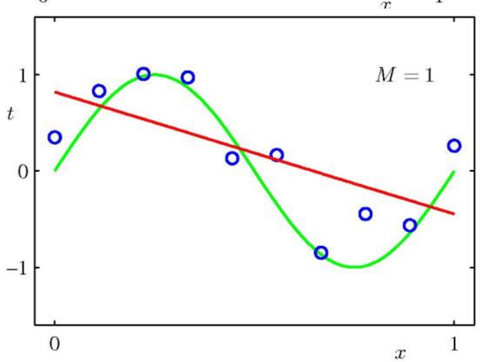
$$y = P(x) = \sum_{k=1}^{p} a_k x^k + b$$

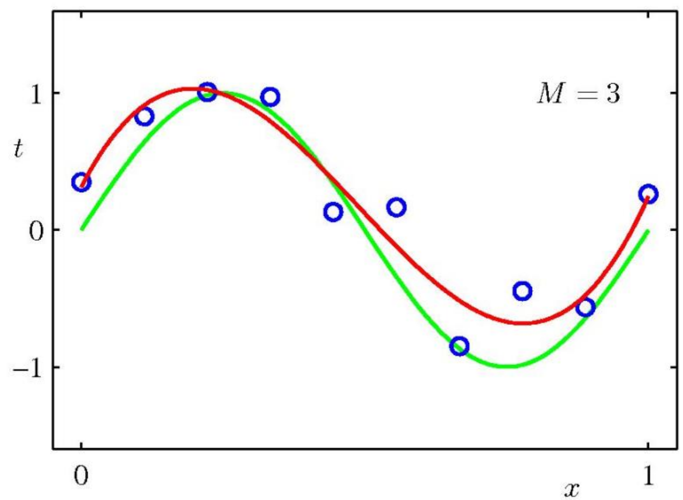
- p is the degree of the polynomial
- a_k and b are the trainable coefficients

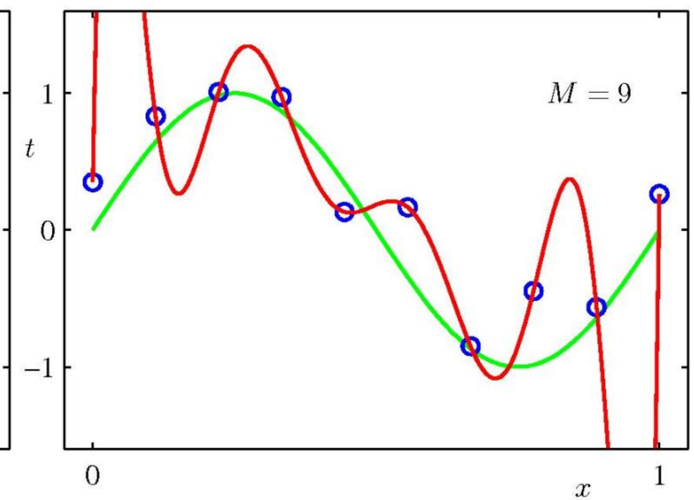
Polynomials can perfectly interpolate a set of points

- A set of data consisting of m points can be perfectly interpolated with a polynomial of degree m-1
  - 2 points define a unique line, 3 points define a unique parabola (or a line, if aligned) and so on…
  - Increasing the degree of the polynomial corresponds to decreasing the error.









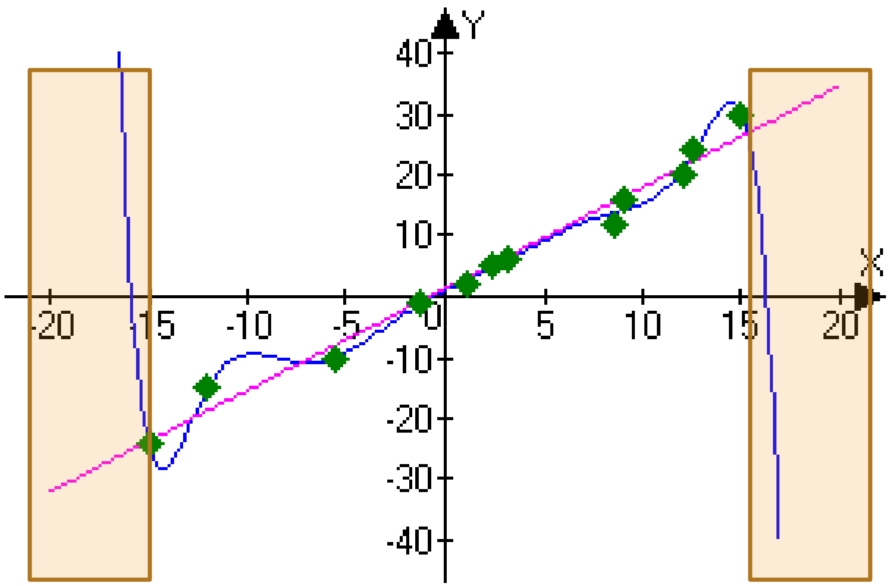
Bishop C, Pattern recognition and Machine Learning, Springer

values for parameters

	M=0	M=1	M=3	M=9
b	0.19	0.82	0.31	0.35
a_1		-1.27	7.99	232.37
a_2			-25.43	-5321
a_3			17.37	48568
a_4				-231639
a_5				640042
a_6				-1061800
a_7				1042400
a_8				-557682
a_9				125201

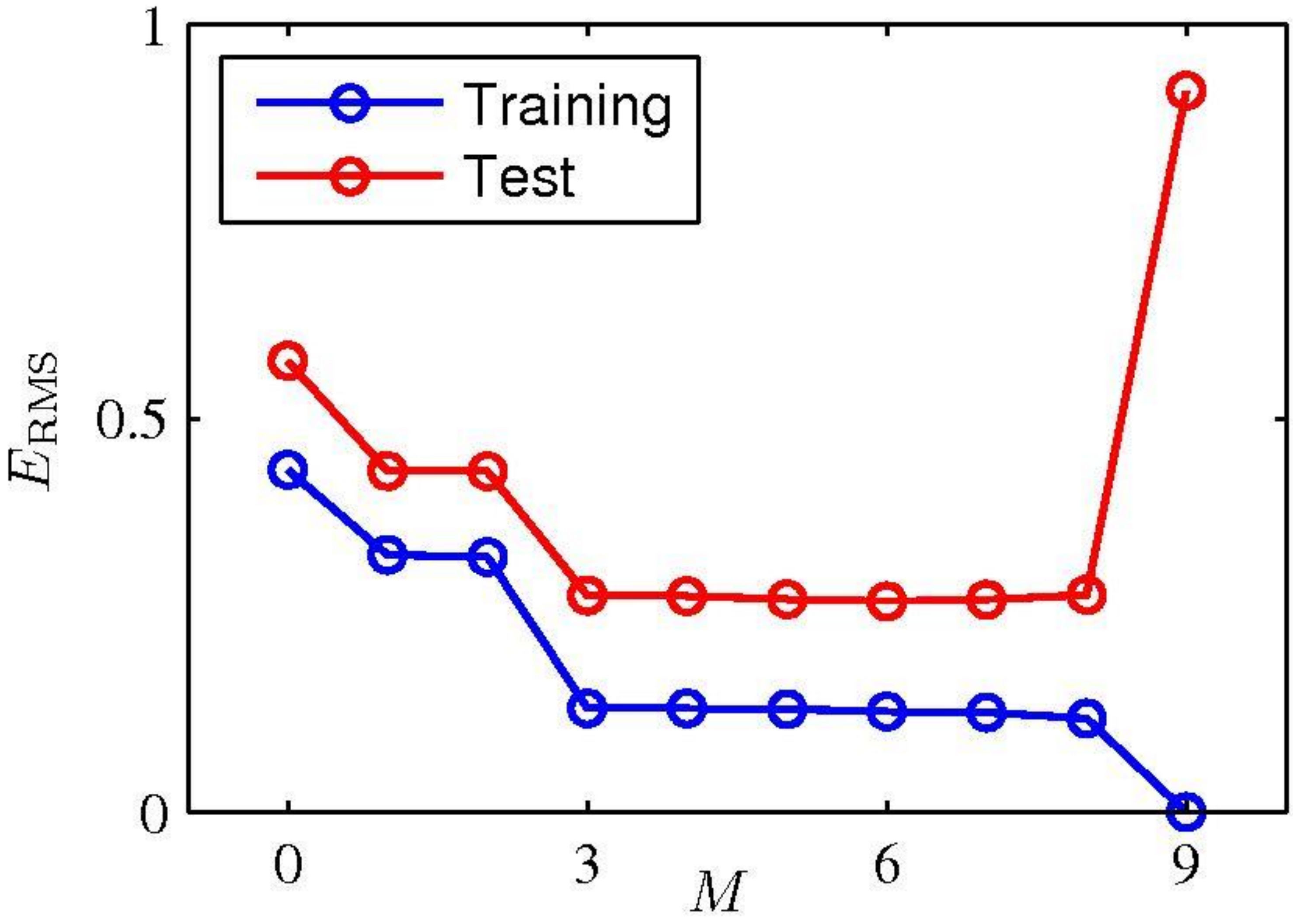
But…

- The interpolation is useless and do not gives a good predictive model for extrapolation→OVERFITTING

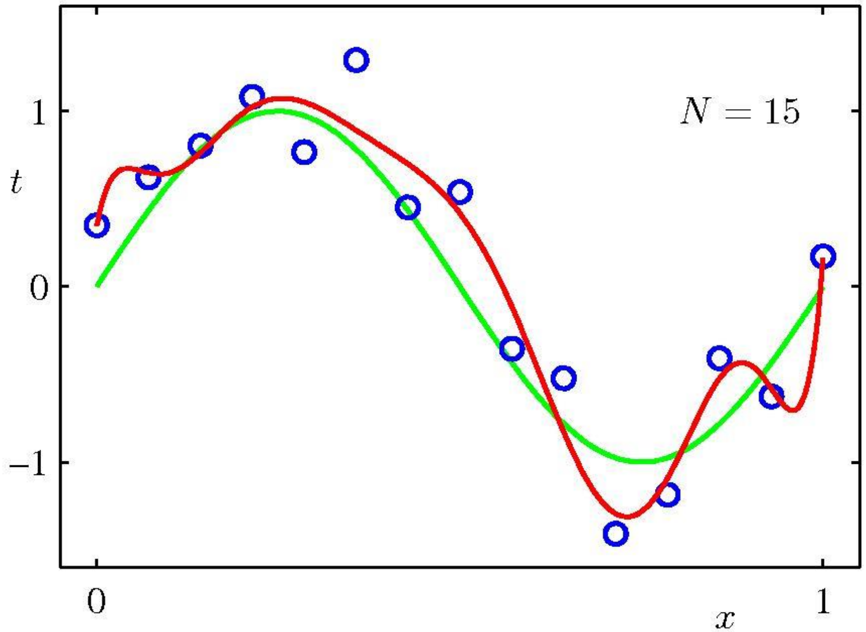


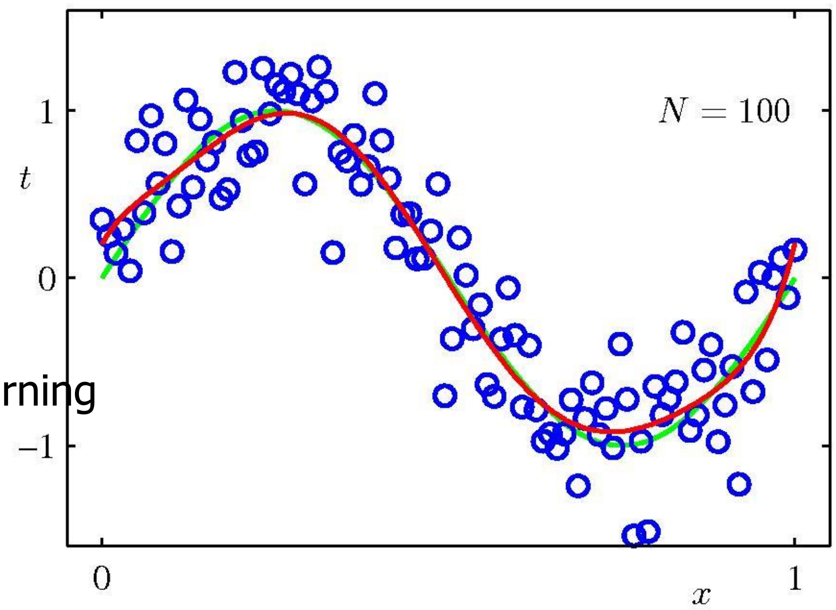
We cannot use the model for predicting Y in this region

Overfitting



Low number of points increases risk of overfitting





Bishop C Pattern Recognition and Machine Learning

High values for parameters

	M=0	M=1	M=3	M=9
b	0.19	0.82	0.31	0.35
a_1		-1.27	7.99	232.37
a_2			-25.43	-5321
a_3			17.37	48568
a_4				-231639
a_5				640042
a_6				-1061800
a_7				1042400
a_8				-557682
a_9				125201

Bishop C Pattern Recognition and Machine Learning

Regularized Error function

Discouraging high values for coefficients

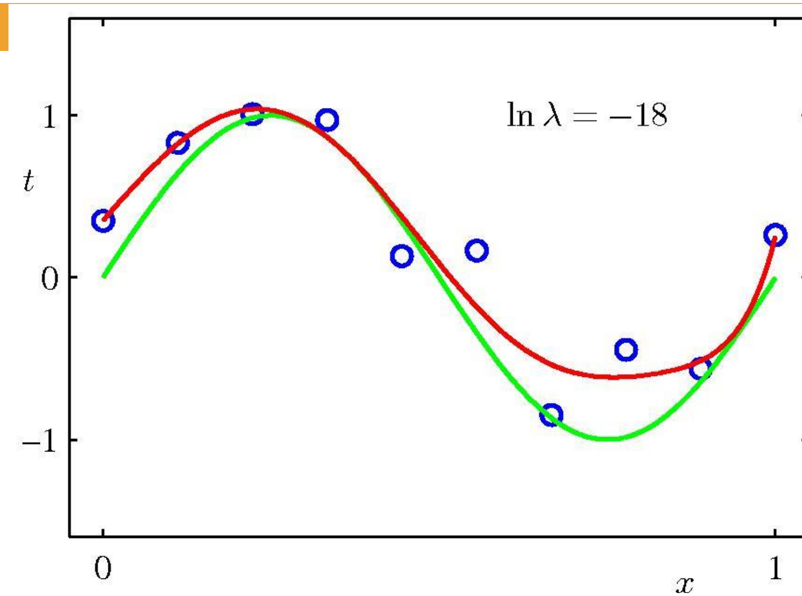
$$E = \sum_{i=1}^{m}\left[y_i - P(x_i)\right]^2 + \lambda \sum_{k=1}^{p} a_k{}^2$$

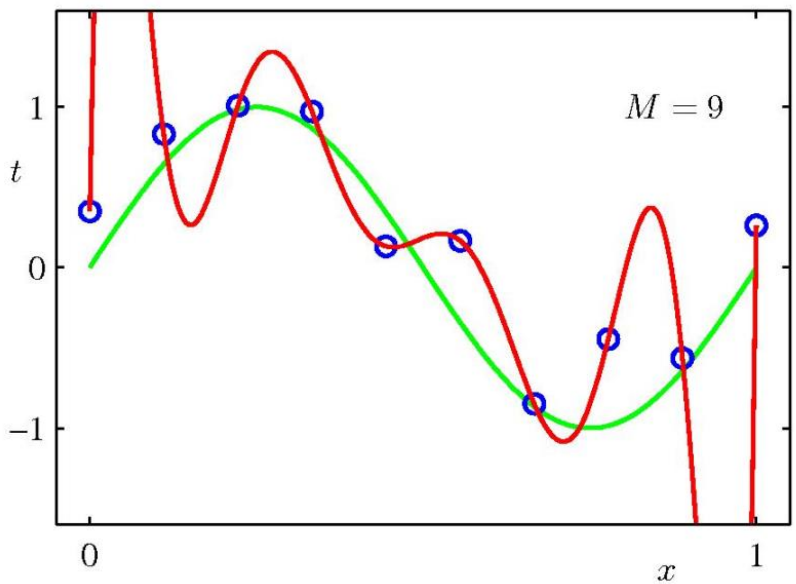
Parameter weighting the strength of regularization

Values for parameters for M=9

	ln λ=0	ln λ=-18	ln λ=-∞
w_0	0.13	0.35	0.35
w_1	-0.05	4.74	232.37
w_2	-0.06	-0.77	-5321
w_3	-0.05	-31.97	48568
w_4	-0.03	-3.89	-231639
w_5	-0.02	55.28	640042
w_6	-0.01	41.32	-1061800
w_7	0.00	-45.95	1042400
w_8	0.00	-91.53	-557682
w_9	0.01	72.68	125201

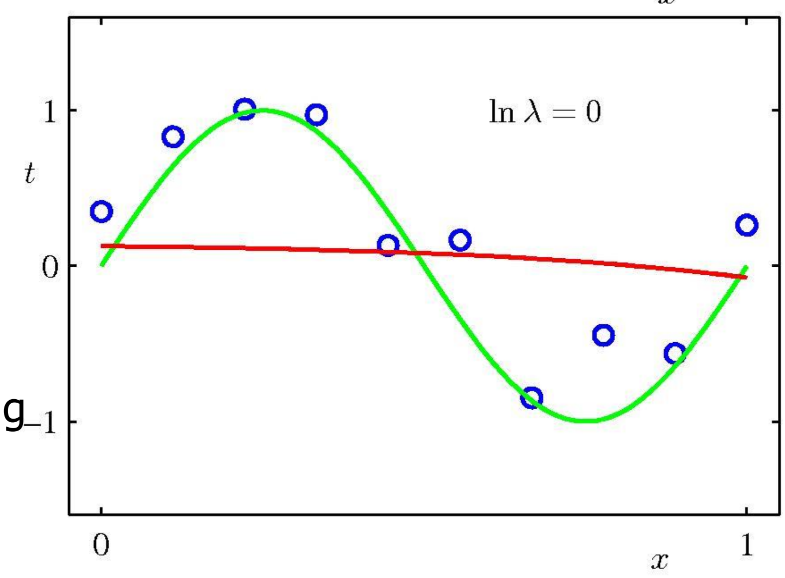
Bishop C Pattern Recognition and Machine Learning



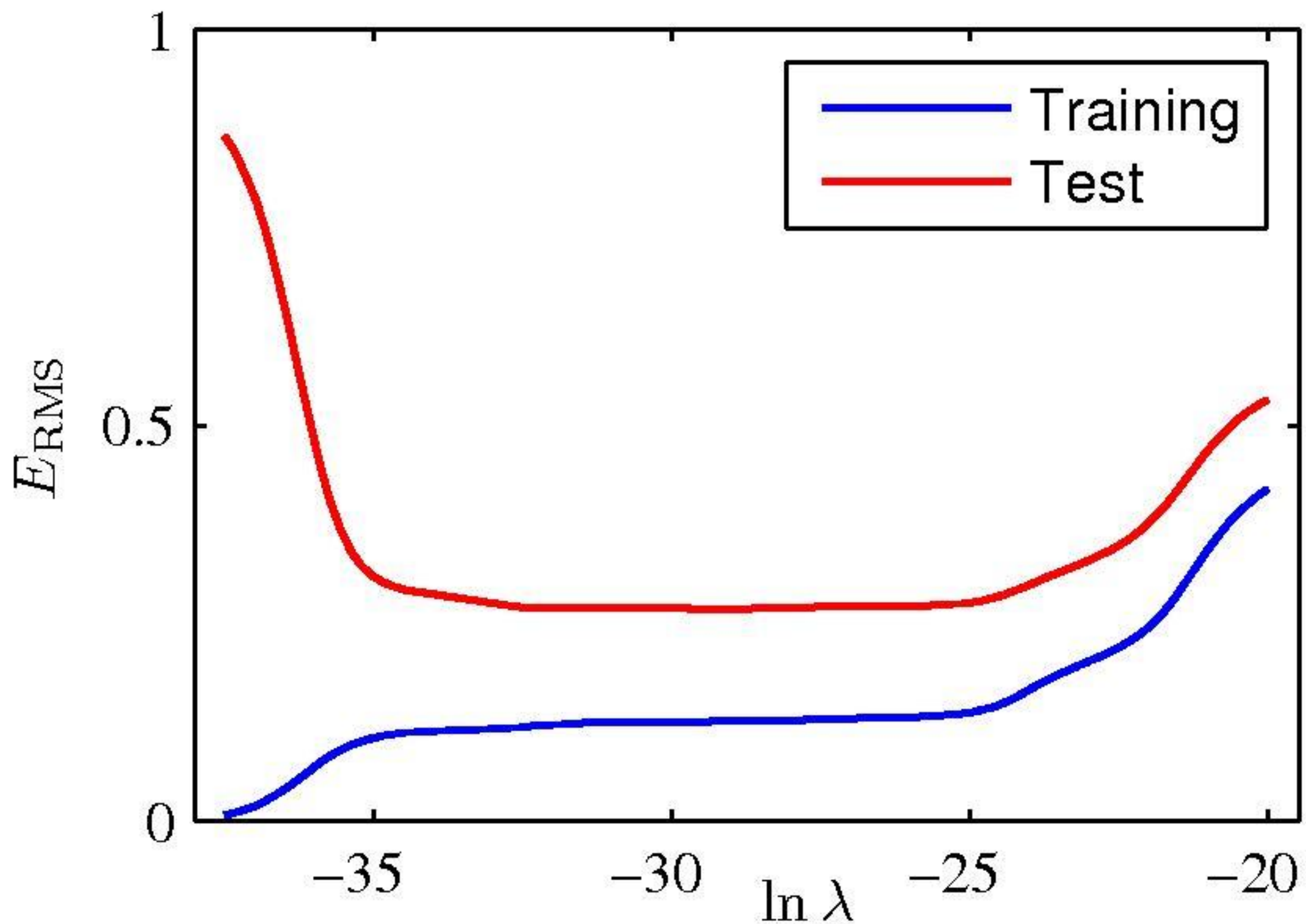


λ=0

Bishop C Pattern Recognition and Machine Learning



Regularized regression avoids overfitting



Bishop C Pattern Recognition and Machine Learning