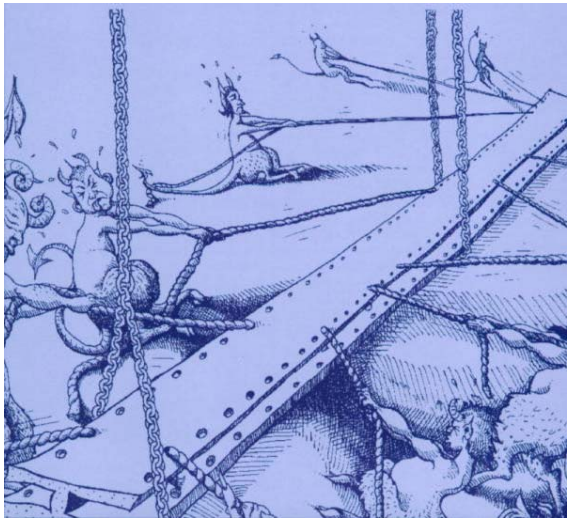
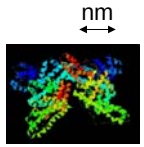
Regression und Korrelation



regression:
Zurückführung,
Rückschreiten

correlation:
Wechselbeziehung

Praktische Annäherung (Beispiel1)



1 St. HSA Molekül

wieviele Eiweissmoleküle sind in dem Blutplasma? (Stück, mol, g, …)

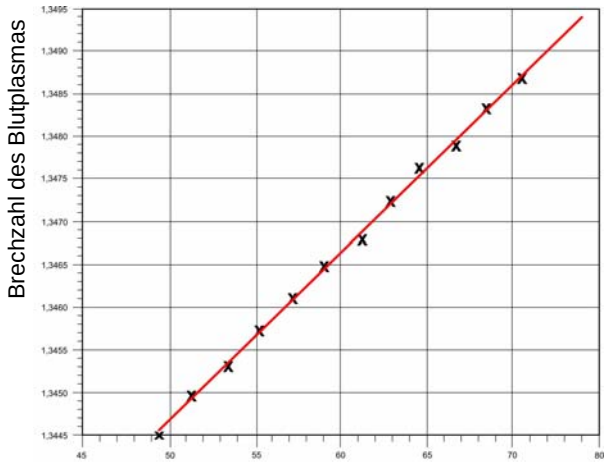
wie gross ist die Eiweisskonzentration des Blutplasmas? (St/L, mol/L, g/L)

bei Patienten in Nephrose (schwere Nierenkrankheit) nimmt der Wert stark ab

direkte Methode:
Bestimmung der Anzahl der Moleküle in einem Volumen(?)

indirekte Methode :
mit Hilfe einer (einfach) messbaren physikalischen Grösse, die steht in streng monoton wachsendem Zusammenhang zu der unbekannten Grösse
(die solche einfachste Funktion ist ...)

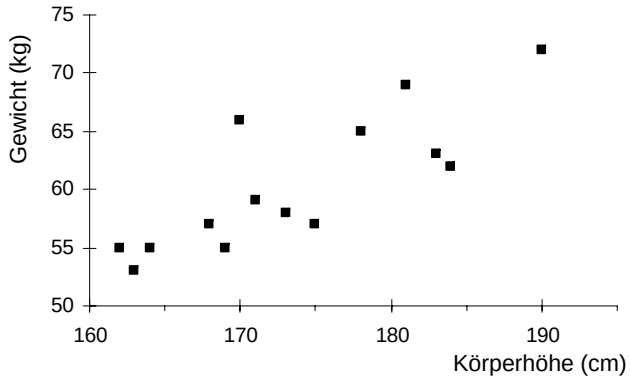
Bemerkung:
das Licht breitet sich in Blutplasma langsamer, wenn die Plasmaeiweisskonzentration grösser ist, d.h. das Licht hat grössere Brechzahl (deterministischer Zusammenhang, Messfehler)



Plasma-
eiweiss-
konzentration
(g/L)

(Beispiel2)

Daten aus einer Studentengruppe (Sept. 1994) (zusammengehörige Wertepaare)

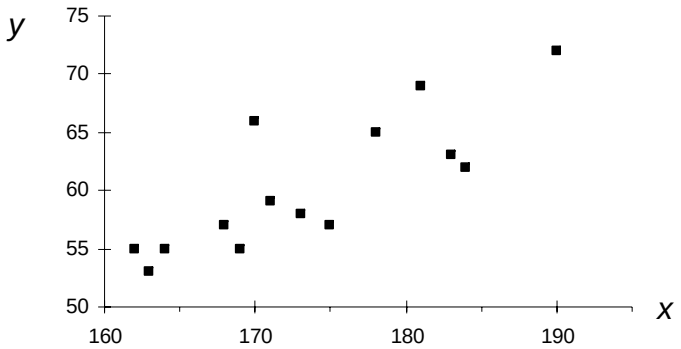


cm	kg
162	55
163	53
164	55
168	57
169	55
170	66
171	59
173	58
175	57
178	65
181	69
183	63
184	62
190	72

was für eine Tendenz kann man bemerken?

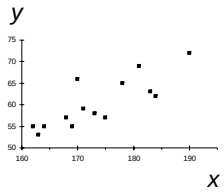
Die **Korrelationsrechnung** beschäftigt sich mit dem **symmetrischen** Zusammenhang zweier Zufallsgrössen

positive Korrelation: je mehr, desto mehr
negative Korrelation:je mehr, desto weniger



hier: positive Korrelation

Regressionsannäherung



Sucht man einen Funktionszusammenhang zwischen einer (oder mehreren) **unabhängigen Variable** (x) und einer abhängigen Variable (y)

Voraussetzungen: x und y numerische und stetige Merkmale, y Zufallsgrösse (ihre Grösse wird nicht nur von der unabhängigen Variable, sondern durch den Zufall beeinflusst)

Regressionsmodell fixiert den Typ der Funktion:

lineare F.	$y = (ax + b) + h$ (a: Steigung, b: Achsenabschnitt)
polinomiale F.	$y = a + b_1x + b_2x^2 + ... + b_nx^n + h$
exponentiale F.	$y = ab^x\ h$ oder $y=ab^x + h$
Potenzfunktion	$y = ax^b\ h$ oder $y=ax^b + h$

und **wie wirkt der Zufall** auf die abhängige Variable
additiver Fehler ($+ h$) oder **multiplikativer** Fehler ($\cdot h$)

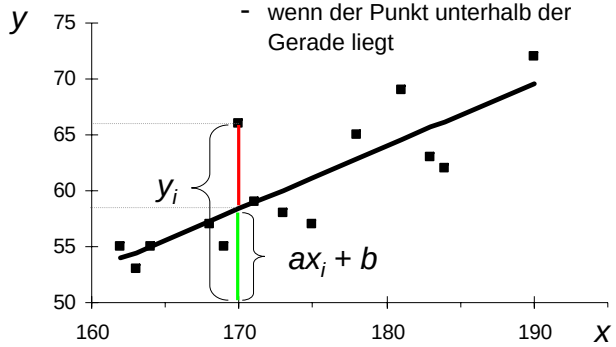
Das einfachste Regressionsmodell: **lineare Regression**

lineare Funktion: $y = (ax + b) + h$

$h_i = y_i - (ax_i + b)$

\+ wenn der Punkt (x_i, y_i) oberhalb der Gerade liegt

\- wenn der Punkt unterhalb der Gerade liegt

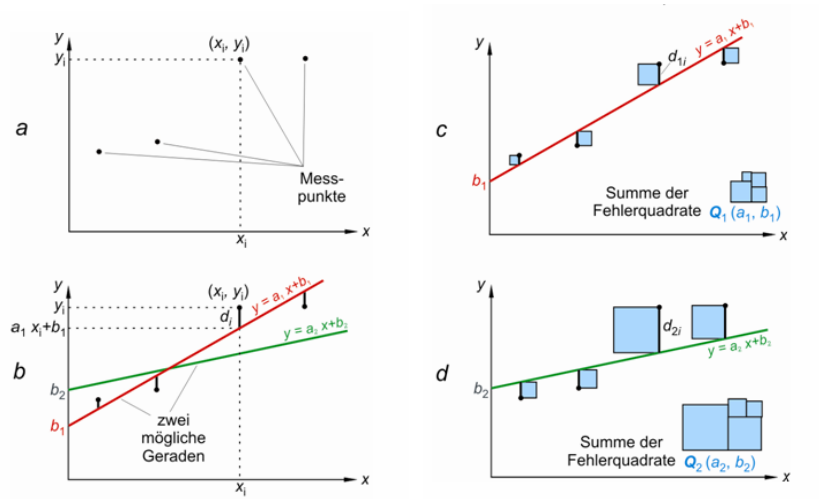


	x_i	y_i
1	162	55
2	163	53
3	164	55
4	168	57
5	169	55
6	170	66
7	171	59
8	173	58
9	175	57
10	178	65
11	181	69
12	183	63
13	184	62
14	190	72

Beste Gerade: Summe der Fehlerquadrate ist minimal (Methode der kleinsten Quadraten)

Suche nach der Geraden ($y = ax + b$) mit bester Näherung der Messpunkte

a: Steigung
b: Achsenabschnitt



Pr.Buch Abb. 13

die (quadratische) **Fehlerfunktion**:

$$Q(...) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2$$

unabhängige Variablen?

a und *b*

$Q(a,b)$!

Funktionszusammenhang für *a* und *b*?

quadratische Zusammenhänge

Wie sehen diese Funktionen aus?

Parabeln mit unterschiedlicher Öffnung

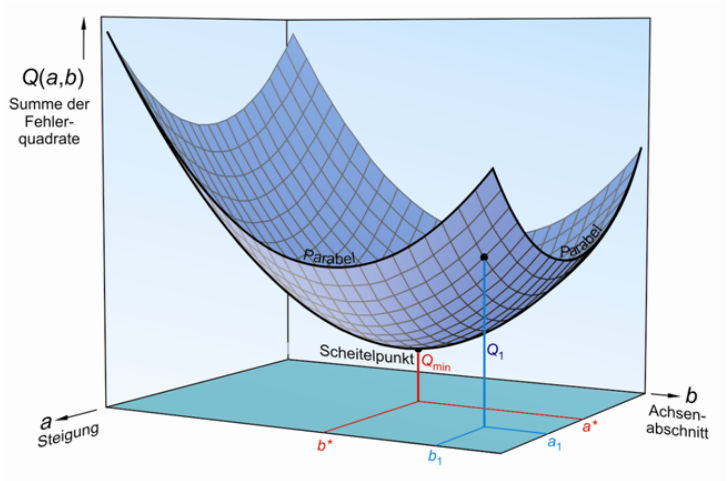
Besitzen diese Funktionen Maxima oder Minima?

die Graphen sind oben geöffnete Parabeln mit Minima

Lineare Regression

$$Q(a, b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2$$

Fehlerfunktion



Pr.Buch Abb. 14

Lineare Regression

$$Q(a, b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2 = \min.$$

Minimalisierung der Fehlerfunktion

Möglichkeiten:

1. quadratische Ergänzung

   z.B. $y = x^2-6x+14 = (x-3)^2+5$, Minimum: $x = 3$

2. Differentialrechnung

   Differentialquotient: Steigung der Tangente

   an dem Minimum/Maximum der Kurve ist die Steigung der Tangente gleich null

   2 Gleichungen, 2 Unbekannten

„Die beste" Steigung: ($y = ax + b$)

$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

oder $a^* = \frac{s_{xy}^2}{s_x^2}$

„Der beste" Achsenabschnitt:

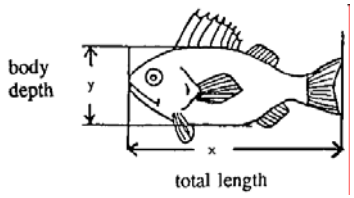
$$b^* = \bar{y} - a^* \cdot \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - a^* \frac{\sum_{i=1}^{n} x_i}{n}$$

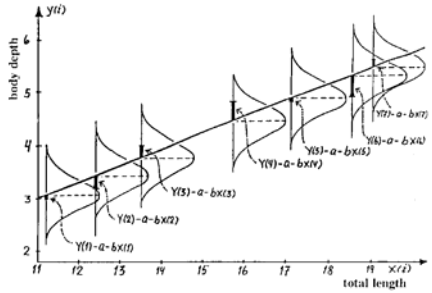
wo $s_{xy}^2 = \frac{Q_{xy}}{n-1}$: Kovarianz

Bedingungen zur Anwendung

(Unter welchen Bedingungen kann man eine lineare Regression durchführen?)

1. Es gibt eine lineare Korrelation zwischen *x* und *y*.
2. Die Messpunkte in einer Stichprobe sind unabhängige Messpunkte.
3. Für alle fixierte *x*-Werte ist die Verteilung von *y* normal.
4. Die Verteilung von *y* für alle *x*-Werte hat dieselbe Varianz.
5. Man kann die *x*-Werte ohne Fehler messen.

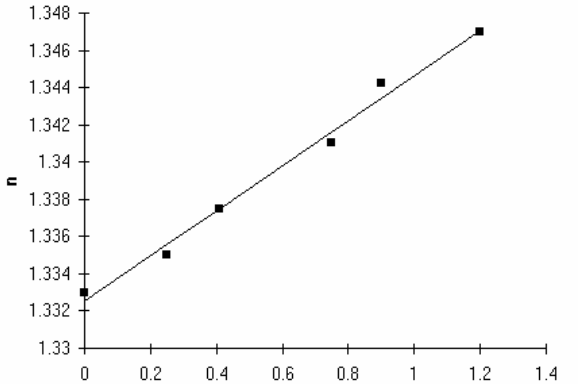




http://www.fao.org/docrep/w5449e/w5449e04.htm

Beispiel: Refraktometrie

	A	B	C
1	unbekannte	Konzentration	Brechzahl
8	i	c (mol/l)	n
9		0	1.333
10		0.25	1.3350
11		0.41	1.3375
12		0.75	1.3410
13		0.90	1.3442
14		1.20	1.3470
15	4	0.19	1.3348
16	10	1.13	1.3462
17	n=ac+n0		
18	n0=	1.3325	
19	a=	0.0121	
20	r=	0.995	



Wie gut passen die Messpunkte an die Regressionsgerade?

Korrelationsrechnung beschreibt die lineare Beziehung zwischen zwei oder mehr statistischen Variablen

es beschreibt die Stärke der Korrelation
es gibt starke und schwache Korrelation

Korrelationskoeffizient
(Pearson)

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}$$

der Zähler ist gleich dem Zähler der Steigung der Regressionsgerade (der Nenner ist in beiden Fällen positiv)

$$a^* = \frac{Q_{xy}}{Q_{xx}}$$ ⇒ positive Steigung: *r* ist positive Zahl

negative Steigung: *r* ist negative ahl

$$-1 \le r \le 1$$

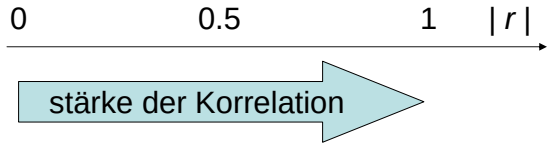
weitere Bemerkungen:

$$-1 \le r \le 1$$

Korrelationskoeffizient
(Pearson)

$$0 \le r^2 \le 1$$

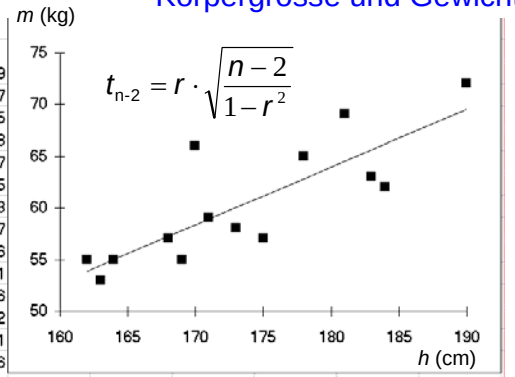
Bestimmtheitsmass
(coefficient of determination)



t-Test zur Korrelationsanalyse

Gibt es eine Beziehung zw. der Körpergrösse und Gewicht?

Körperhöhe (cm)	Gewicht (kg)		
162	55	53.929	
163	53	54.487	
164	55	55.045	
168	57	57.278	
169	55	57.837	
170	66	58.395	
171	59	58.953	
173	58	60.07	
175	57	61.186	
178	65	62.861	
181	69	64.536	
183	63	65.652	
184	62	66.211	
190	72	69.56	
	m=	0.5583	-36.50955 =b
		0.1131	19.66358
r=	0.818505	0.6699	3.492297
n=	14	24.358	12
t=	6.030	297.07	146.3537

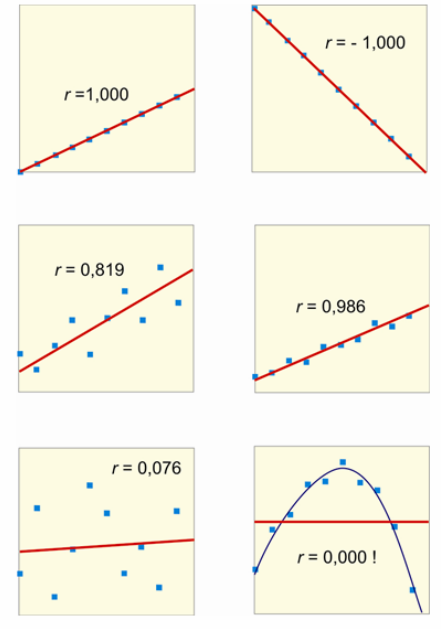


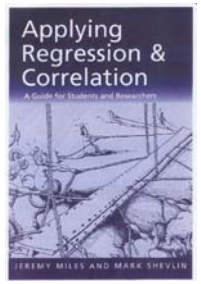
$$t_{n-2} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

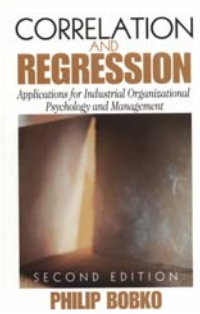
$|t| = 6.030 > t_{12,\text{krit}(0,05)} = 2.179 \Rightarrow H_0$ ist falsch (p<0.05)

$|t| = 6.030 > t_{12,\text{krit}(0,01)} = 3.055 \Rightarrow H_0$ ist falsch (p<0.01)

Beispiele:

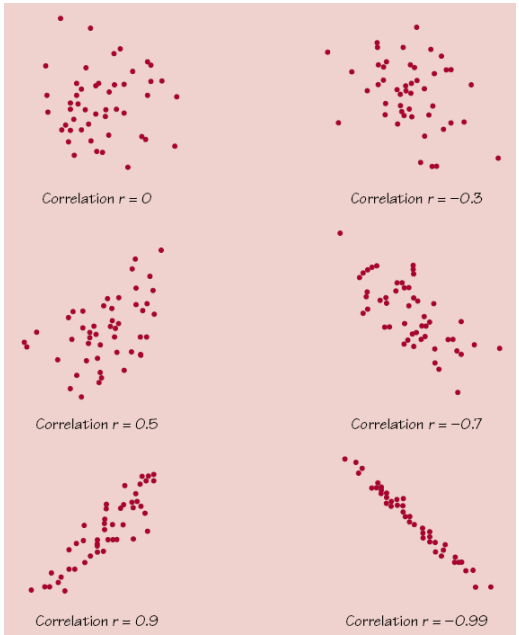




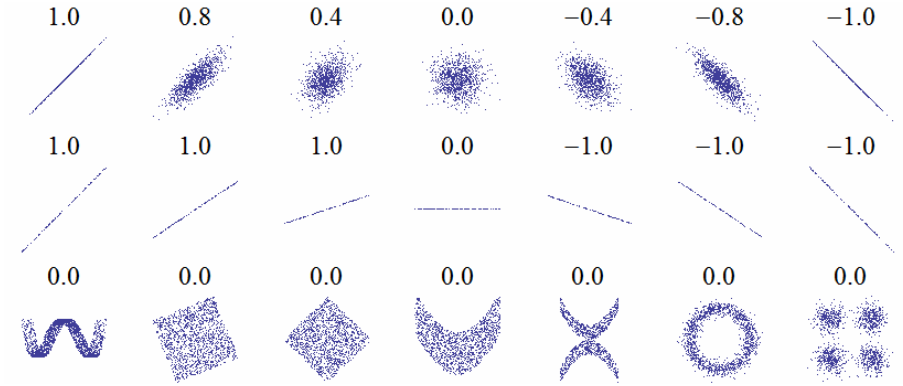


Pr.Buch Abb. 15

Punkt-diagrammen

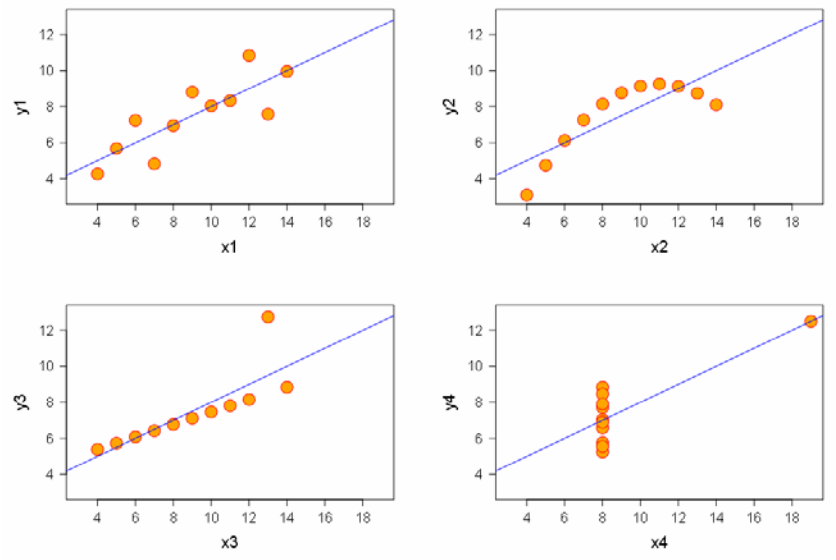


http://www.cshgreenwich.org/faculty/byrne/chapter%203.pdf



http://en.wikipedia.org/wiki/Image:Correlation_examples.png

Extrembeispiel: r=0.816, y = 3 + 0.5x (Anscombe's quartet)

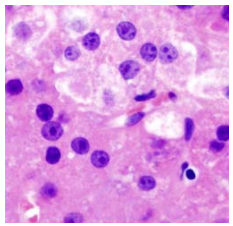


http://en.wikipedia.org/wiki/Anscombe%27s_quartet

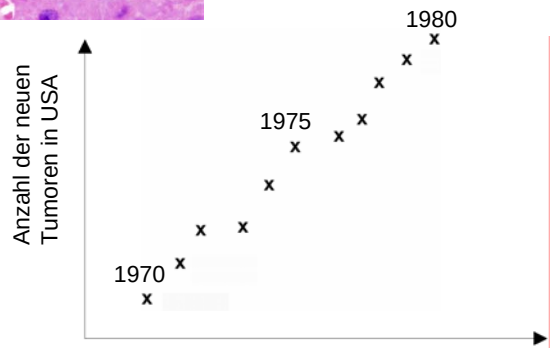
Anscombe's quartet comprises four datasets which have identical simple statistical properties, yet which are revealed to be very different when inspected graphically. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician F.J. Anscombe to demonstrate the importance of graphing data before analyzing it, and of the effect of outliers on the statistical properties of a dataset.

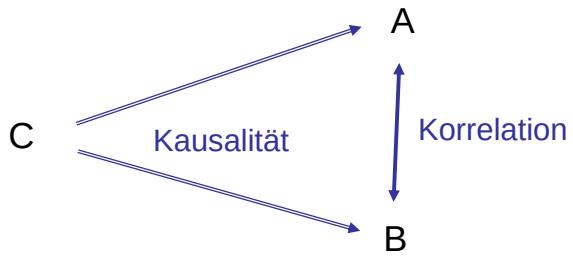
Property	Value
Mean of each x variable	9.0
Variance of each x variable	10.0
Mean of each y variable	7.5
Variance of each y variable	3.75
Correlation between each x and y variable	0.816
Linear regression line	y = 3 + 0.5x

http://en.wikipedia.org/wiki/Anscombe%27s_quartet



Korreliert heisst **nicht** notwendigerweise **kausal verknüpft**(!)





Diagnostik - Therapie

Wichtig:

- Die „beste Gerade“ hat die kleinste Quadratische Fehlersumme.
- Graphische Darstellung ist notwendig.
- Berechnung des Korrelationskoeffizienten!
- Korrelations *t*-Test
- Korrelation ist nicht unbedingt Kausalität

Wie kann man nichtlineare Zusammenhänge Auswerten?

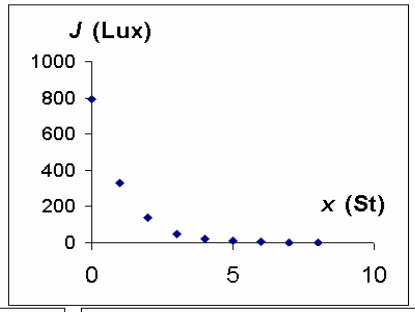
Beispiele für nichtlineare Funktionen:

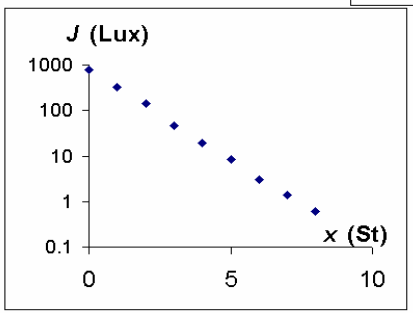
polinomiale F.	$y = a + b_1x + b_2x^2 + ... + b_nx^n + h$
exponentiale F.	$y = ab^x\ h$ oder $y=ab^x + h$
Potenzfunktion	$y = ax^b\ h$ oder $y=ax^b + h$

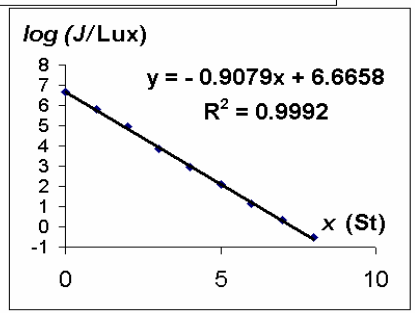
1. Zurückführung auf lineare Regression mit Transformation.
2. Direkte Anwendung des Prinzips der kleinsten Quadrate

Zusammenhang zwischen der Intensität und der Schichtdicke

x St	J (Lux)	ln (J/lux)
0	790	6.6720
1	330	5.7991
2	140	4.9416
3	46.9	3.8480
4	19.6	2.9755
5	8.3	2.1163
6	3.1	1.1314
7	1.4	0.3365
8	0.6	-0.5108







Direkte Verwendung der Methode der kleinsten Quadrate

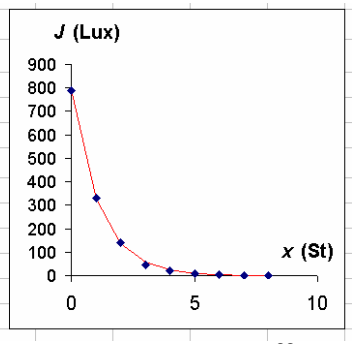
Die quadratische Fehlersumme muss berechnet und minimalisiert werden so dass die zu anpassende Parametern variiert werden. (Z.B. in Excel: „Solver“)

x St	J (Lux)	$J_{expfunct}$ (Lux)	Quadr.Abw. (Lux^2)
0	790	790.98	0.96060
1	330	327.54	6.04591
2	140	135.63	19.06842
3	46.9	56.165	85.84204
4	19.6	23.258	13.37880
5	8.3	9.6309	1.77131
6	3.1	3.9881	0.78875
7	1.4	1.6515	0.06323
8	0.6	0.6839	0.00703
			127.926

Angepasste Funtion: $J=J_0\ e^{-mx}$

Parametern der Anpassung

J_0=	790.98
μ=	0.8817



Quadratische Fehlersumme