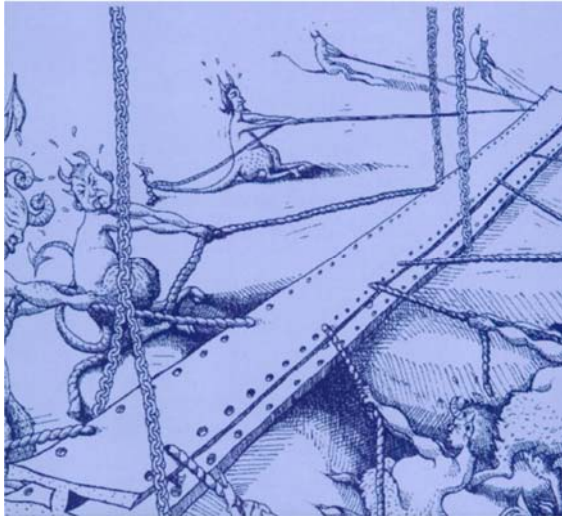


## Regression und Korrelation



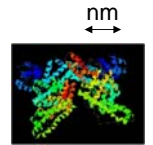
KAD 2017.11.22

**regression:**  
Zurückführung,  
Rückschreiten

**correlation:**  
Wechselbeziehung

## Praktische Annäherung (Beispiel1)

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



1 St. HSA Molekül

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:

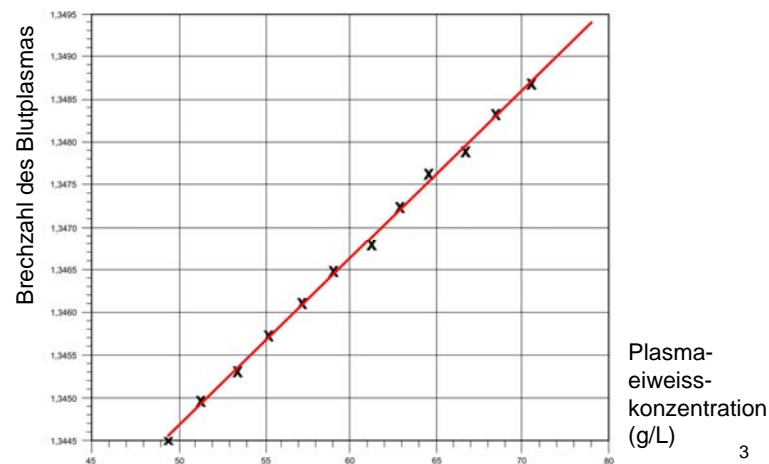
Bestimmung der Anzahl der Moleküle in einem Volumen(?)

### 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 ...)

2

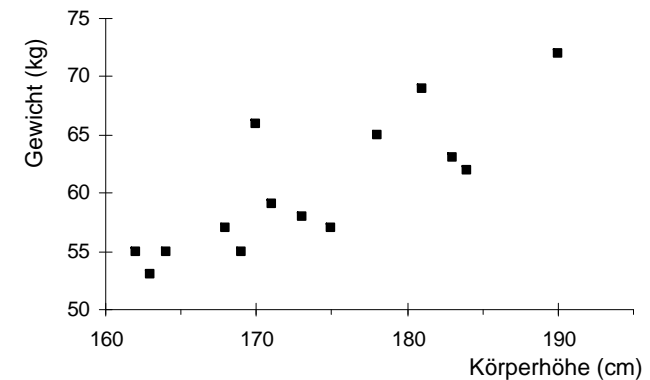
Bemerkung:  
das Licht breitet sich in Blutplasma langsamer, wenn die  
Plasmaeiweisskonzentration grösser ist, d.h. das Licht hat  
grössere Brechzahl (deterministischer Zusammenhang, Messfehler)



3

## (Beispiel2)

Daten aus einer Studentengruppe E2  
(Sept. 1994) (zusammengehörige  
Wertepaare)



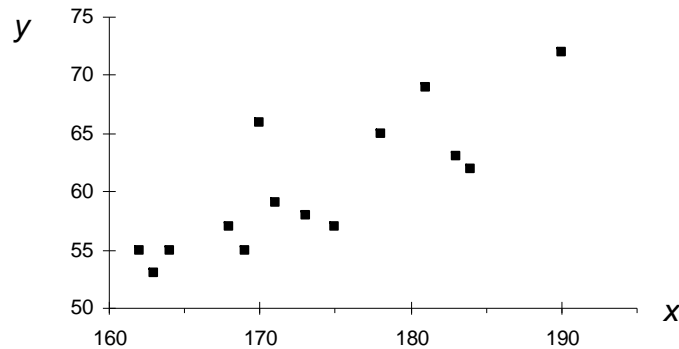
cm	kg
162	55
163	53
164	55
168	57
169	55
170	66
171	59
173	58
175	57
178	65
181	69
183	63
184	62
190	72

was für eine Tendenz kann man bemerken?

4

## Die Korrelationsrechnung beschäftigt sich mit dem symmetrischen Zusammenhang zweier Zufallsgrößen

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

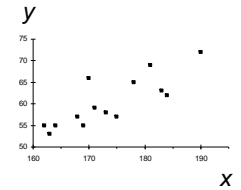


hier: positive Korrelation

5

## 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.  $y = a + b_1x + b_2x^2 + \dots + b_nx^n + h$

exponentiale F.  $y = ab^x h$

Potenzfunktion  $y = ax^b h$

und **wie wirkt der Zufall** auf die abhängige Variable

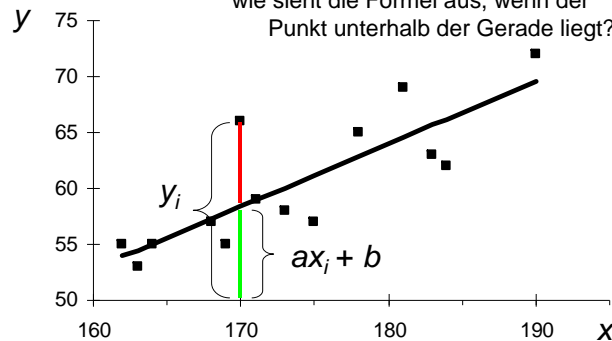
**additiver Fehler (+ h)** oder **multiplikativer Fehler (· h)**

6

## 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  
wie sieht die Formel aus, wenn der Punkt unterhalb der Gerade liegt?



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

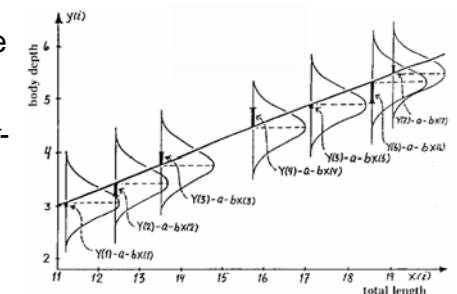
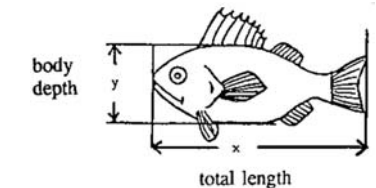
	$x_i$	$y_i$
1	162	55
2	163	53
3	164	55
4	168	57
5	169	55
6	170	66
7	171	59
8	173	58
9	175	57
10	178	65
11	181	69
12	183	63
13	184	62
14	190	72

7

## 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.



8

die (quadratische) **Fehlerfunktion**:

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

unabhängige Variablen?  
*a und b*

Funktionszusammenhang für *a* und *b*?

*quadratische Zusammenhänge*

Wie sehen diese Funktionen aus?

*Parabeln mit unterschiedlicher Öffnung*

Besitzen diese Funktionen Maxima oder Minima?

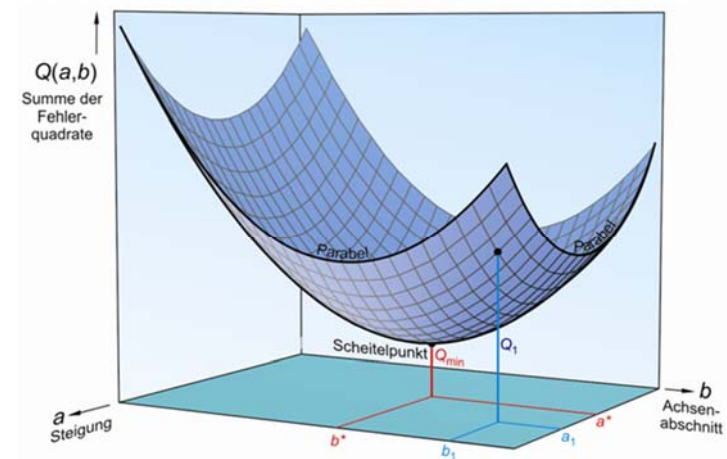
*die Graphen sind oben geöffnete Parabeln mit Minima*

9

## Lineare Regression

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

Fehlerfunktion

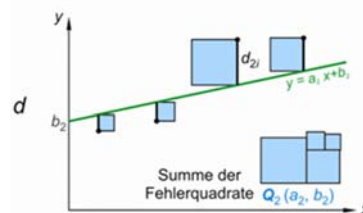
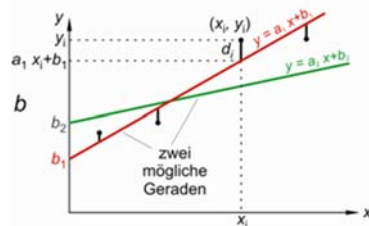
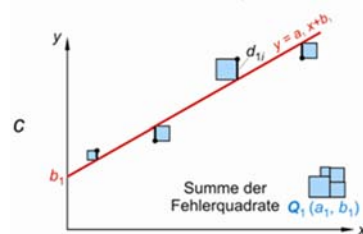
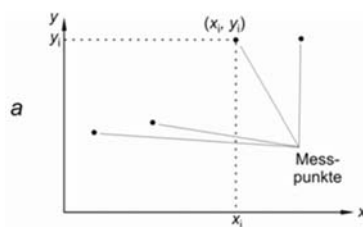


Pr.Buch Abb. 14

10

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

*a*: Steigung  
*b*: Achsenabschnitt



11

Pr.Buch Abb. 13

## Lineare Regression

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

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

12

„Die beste“ Steigung:

$$(y = ax + b)$$

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

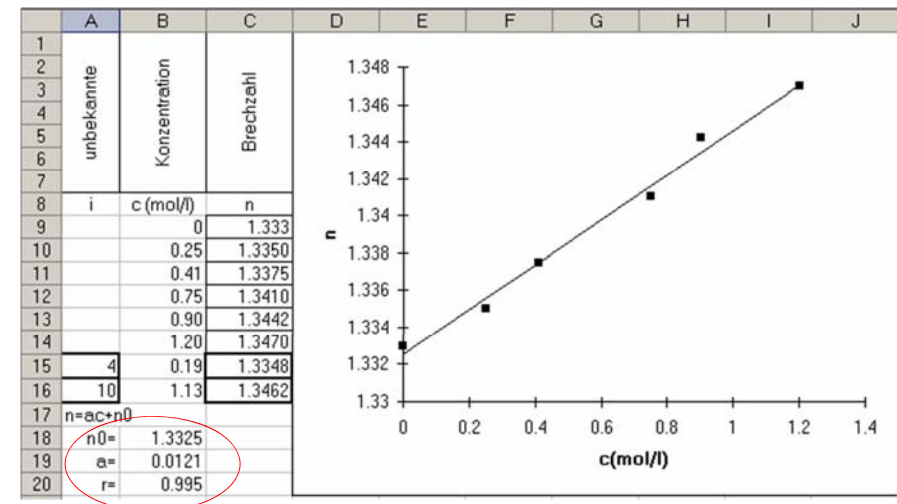
„Der beste“ Achsenabschnitt:

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

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

13

Beispiel: Refraktometrie



14

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 im beiden Fall positiv)

$a^* = \frac{Q_{xy}}{Q_{xx}}$  positive Steigung: r ist positive Zahl  
negative Steigung: r ist negative Zahl  
 $-1 \leq r \leq 1$

15

weitere Bemerkungen:

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

$$0 \leq r^2 \leq 1$$

**Korrelationskoeffizient**  
(Pearson)

**Bestimmtheitsmass**  
(Determinationskoeffizient)

die Variation von y kann durch das Modell erklärte Variation und durch das Modell nicht erklärte Variation zerlegt werden  
z.B.  $r^2 = 0,25$ , das Modell kann erklären 25 % der Variation von y

perfekter linearer Zusammenhang:  $r^2 = 1$ ,  $r = 1$  oder  $-1$   
kein linearer Zusammenhang:  $r^2 = 0$ ,  $r = 0$   
Unabhängigkeit oder nichtlinearer Zusammenhang

16

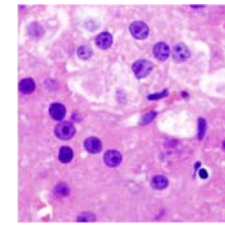
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\frac{\sum x_i y_i}{n} - \left(\frac{\sum x_i}{n}\right)\left(\frac{\sum y_i}{n}\right)}{\sqrt{\left(\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right)\left(\frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2\right)}}$$

aus den Rangwerten berechneter Korrelationskoeffizient:  
**Spearman's Rangkorrelationskoeffizient** ( $r_s$ )  
(Pearson-Korrelation zwischen den Rangwerten)

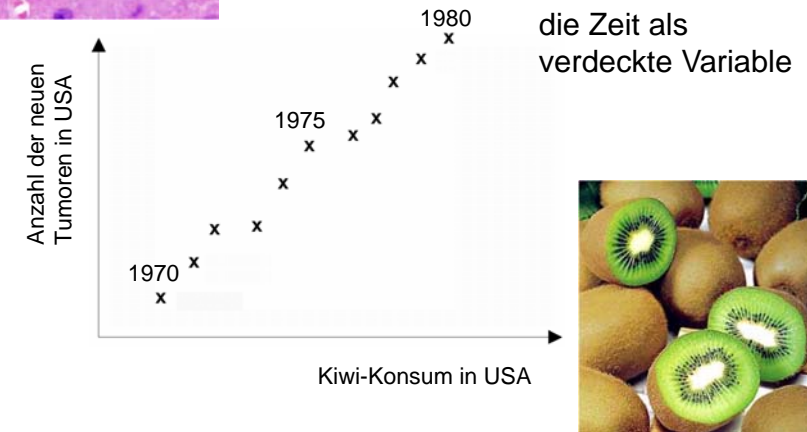
$$r_s = \frac{\sum_{i=1}^n (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R_{x_i} - \bar{R}_x)^2 \sum_{i=1}^n (R_{y_i} - \bar{R}_y)^2}} \quad \tau = \frac{n_c - n_d}{\frac{1}{2} n (n - 1)}$$

noch ein aus den Rangwerten berechneter Korrelationskoeff.:  
**Kendalls Tau** ( $\tau$ )  
(es nutzt nur den Unterschied in den Rängen und nicht die Differenz der Ränge)

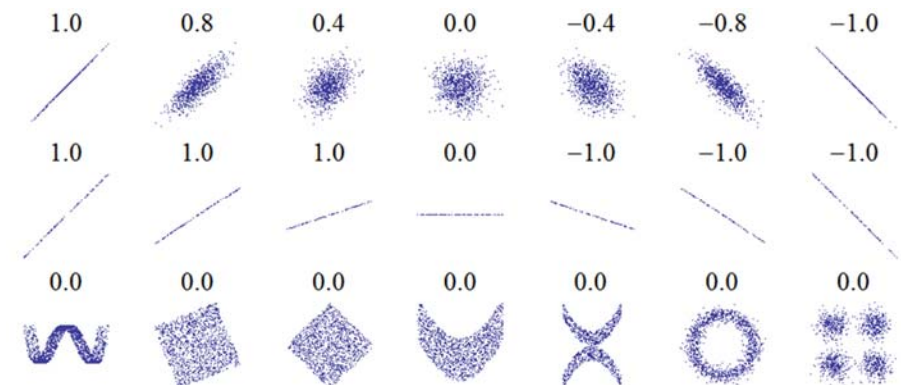
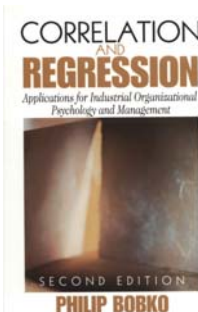
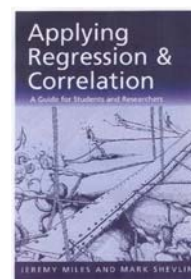
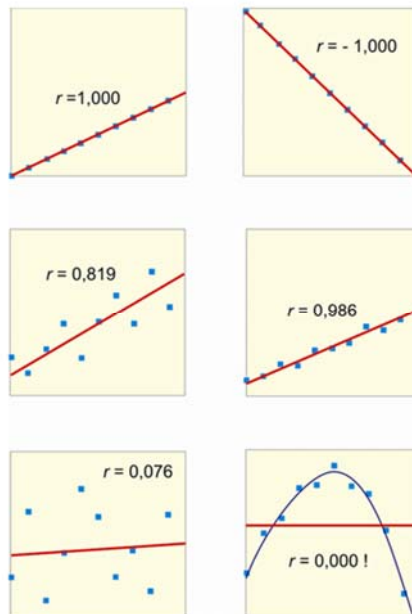
17



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

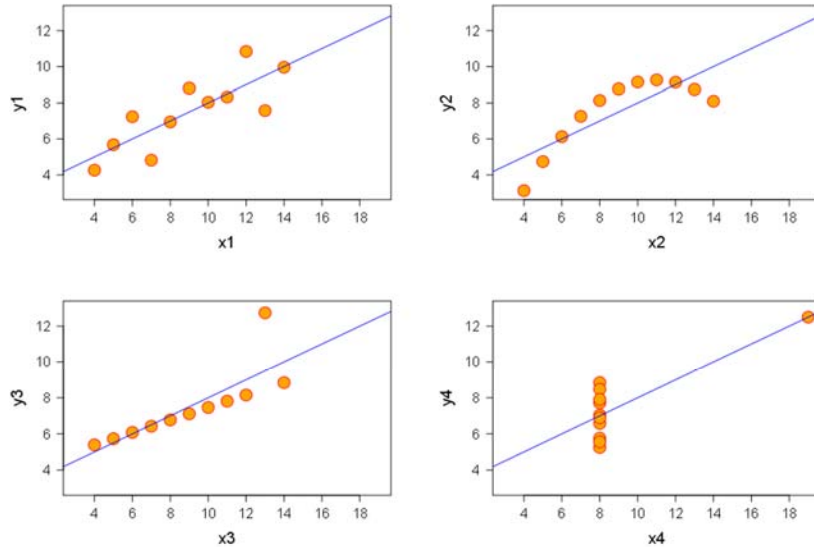


Beispiele:





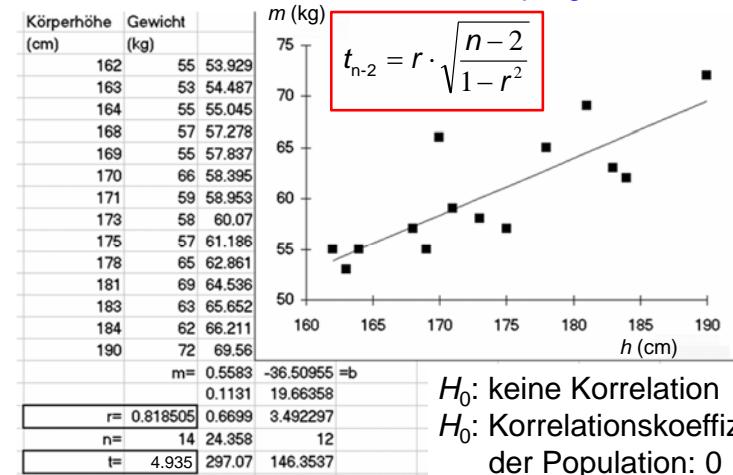
Extrembeispiel:  $r=0.816$ ,  $y = 3 + 0.5x$  (Anscombe's quartet)



[http://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](http://en.wikipedia.org/wiki/Anscombe%27s_quartet)

21

**t-Test zur Korrelationsanalyse** Gibt es eine Beziehung zw. der Körpergröße und Gewicht?



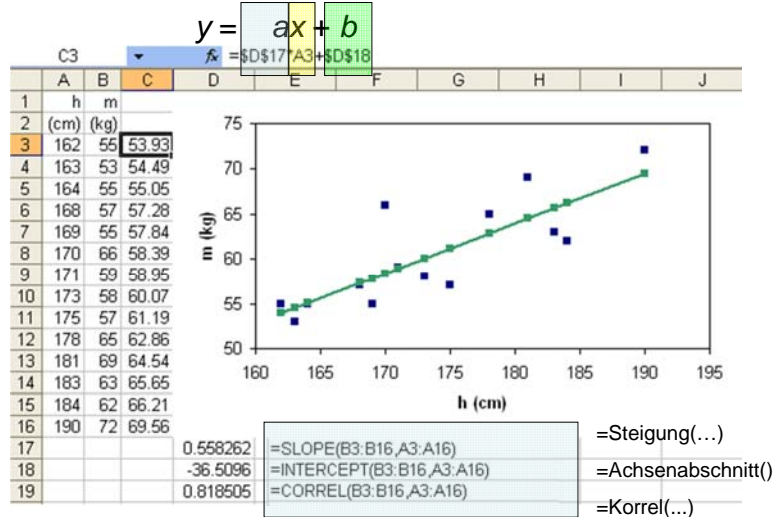
$H_0$ : keine Korrelation  
 $H_0$ : Korrelationskoeffizient der Population: 0

$|t| = 4.935 > t_{12, \text{krit}(0,05)} = 2.179 \Rightarrow H_0$  ist falsch ( $p < 0.05$ )

$|t| = 4.935 > t_{12, \text{krit}(0,01)} = 3.055 \Rightarrow H_0$  ist falsch ( $p < 0.01$ )

22

Steigung, Achsenabschnitt Funktionen in Excel



23

