

# Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

<http://blog.ruediger-braun.net>

Heinrich-Heine-Universität Düsseldorf

11. Dezember 2013

- 1 Datenpaare
  - Korrelation
  
- 2 Lineare Regression
  - Problemstellung
  - Beispiel Bleibelastung
  
- 3 Regression im exponentiellen Modell
  - Beispiel Gendatenbank
  - Exponentielles Modell vs. Lineare Regression
  - Gendatenbank durchgerechnet

# Datenpaare

- Häufig erhebt man zwei Merkmale, um deren Abhängigkeit zu erforschen
- mathematisch stellt man solch ein Ergebnis als Menge von Datenpaaren dar

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

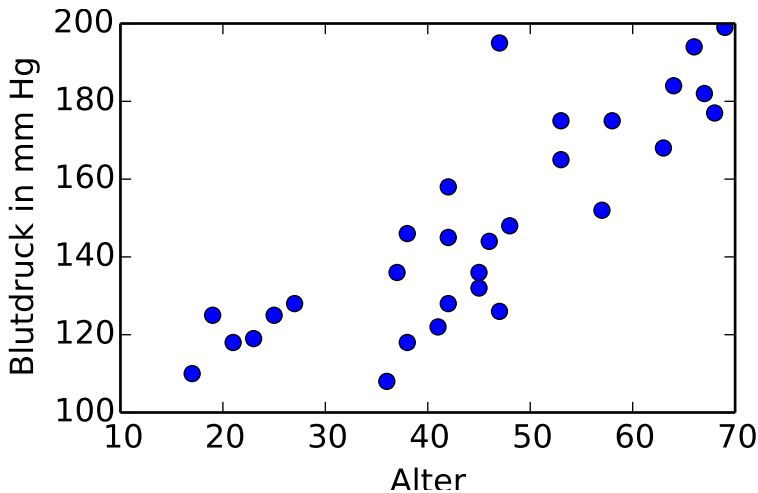
- Beispiel: Zusammenhang von Alter und Blutdruck
- Dann
  - $x_n$ : Alter des  $n$ -ten Probanden in Jahren
  - $y_n$ : Blutdruck des  $n$ -ten Probanden in *mm Hg*

# Blutdruckdaten

Die Tabelle zeigt Alter und Blutdruck von 30 Probanden

(17, 110)	(19, 125)	(21, 118)	(23, 119)	(25, 125)
(27, 128)	(36, 108)	(37, 136)	(38, 118)	(38, 146)
(41, 122)	(42, 128)	(42, 145)	(42, 158)	(45, 132)
(45, 136)	(46, 144)	(47, 126)	(47, 195)	(48, 148)
(53, 165)	(53, 175)	(57, 152)	(58, 175)	(63, 168)
(64, 184)	(66, 194)	(67, 182)	(68, 177)	(69, 199)

## Beispiel: Blutdruckdaten



# Korrelation

- Eine Korrelation zwischen zwei Datensätzen ist eine gemeinsame oder gegenläufige Tendenz.
- Beispielsweise steigt der Blutdruck tendenziell mit dem Alter.
- Gemessen wird die Korrelation durch den empirischen Korrelationskoeffizienten.
- Der empirischen Korrelationskoeffizient beantwortet die Frage  
*Gibt es eine Korrelation?*
- Die Antwort ist “ja”, wenn der empirische Korrelationskoeffizient nahe bei 1 oder bei  $-1$  liegt.

# Empirische Kovarianz

- Empirische Kovarianz: Wie empirische Varianz, aber für Datenpaare.
- Wir haben  $n$  Datenpaare  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- das arithmetische Mittel der  $x_j$  ist  $\bar{x}$ , das der  $y_j$  ist  $\bar{y}$
- die *empirische Kovarianz* von  $x$  und  $y$  ist

$$\begin{aligned}\text{covar}_{\text{emp}}(x, y) \\&= \frac{1}{n-1} \left( (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots \right. \\&\quad \left. + (x_n - \bar{x})(y_n - \bar{y}) \right)\end{aligned}$$

- Formel ohne Pünktchen

$$\text{covar}_{\text{emp}}(x, y) = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{n-1}$$

# Konkrete Berechnung der Kovarianz

- Wir verwenden die erste Spalte der Blutdruckdaten
- $x$  ist das Alter in Jahren,  $y$  der Blutdruck in *mm Hg*
- Man erhält  $\bar{x} = \frac{247 \text{Jahre}}{6} = 41.17 \text{Jahre}$  und  
 $\bar{y} = \frac{845 \text{mm Hg}}{6} = 140.83 \text{mm Hg}$
- Rechnung:

$j$	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$y_j - \bar{y}$	$(y_j - \bar{y})^2$	$(x_j - \bar{x})(y_j - \bar{y})$
1	-25.17	584.2	-30.83	950.5	745.2
2	-14.17	200.8	-12.83	164.6	181.8
3	-0.17	0.0	-18.83	354.6	3.2
4	3.83	14.7	-4.83	23.3	-18.5
5	11.83	139.9	24.17	584.2	286.0
6	22.83	521.2	41.17	1695.0	939.9
$\sum$	-0.02	1460.8	0.02	3772.2	2137.6

- $\text{covar}_{\text{emp}} = \frac{2137.6}{5} = 427.5 \text{Jahre} \times \text{mm Hg}$

# Eigenschaften der empirischen Kovarianz

- $\text{covar}_{\text{emp}}(y, y) = s_y^2$ , wobei  $s_y^2$  die empirische Varianz von  $y$  ist.
- Die empirische Kovarianz skaliert wie die Varianz

# Empirischer Korrelationskoeffizient

- Kennzahl zur Überprüfung gemeinsamer Tendenz
- $s_x$  sei die Stichprobenstreuung der  $x_j$  und  $s_y$  die Stichprobenstreuung der  $y_j$
- dann ist der *empirische Korrelationskoeffizient* gleich

$$r = \frac{\text{covar}_{\text{emp}}(x, y)}{s_x \cdot s_y}$$

- ausgeschrieben bedeutet das

$$r = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\left(\sum_{j=1}^n (x_j - \bar{x})^2\right) \left(\sum_{j=1}^n (y_j - \bar{y})^2\right)}}$$

- Der Korrelationskoeffizient ist dimensionslos

## Fortsetzung: Blutdruckdaten

Weiterhin werden nur die Daten der ersten Spalte zugrunde gelegt.  
Bereits gesehen  $\text{covar}_{\text{emp}} = 427.5 \text{ Jahre} \cdot \text{mm Hg}$

$$s_x^2 = \frac{1460.8 \text{ Jahre}^2}{5} = 292.2 \text{ Jahre}^2$$

$$s_x = \sqrt{292.2 \text{ Jahre}^2} = 17.09 \text{ Jahre}$$

$$s_y^2 = \frac{3772.2 \text{ mm}^2}{5} = 754.4 \text{ mm}^2$$

$$s_y = \sqrt{754.4 \text{ mm}^2} = 27.47 \text{ mm}$$

empirischer Korrelationskoeffizient

$$r = \frac{427.5 \text{ Jahre} \cdot \text{mm Hg}}{17.09 \text{ Jahre} \cdot 27.47 \text{ mm}} = 0.9106$$

# Interpretation

Der Korrelationskoeffizient zeigt an, ob zwei Datensätze eine gemeinsame Tendenz aufweisen

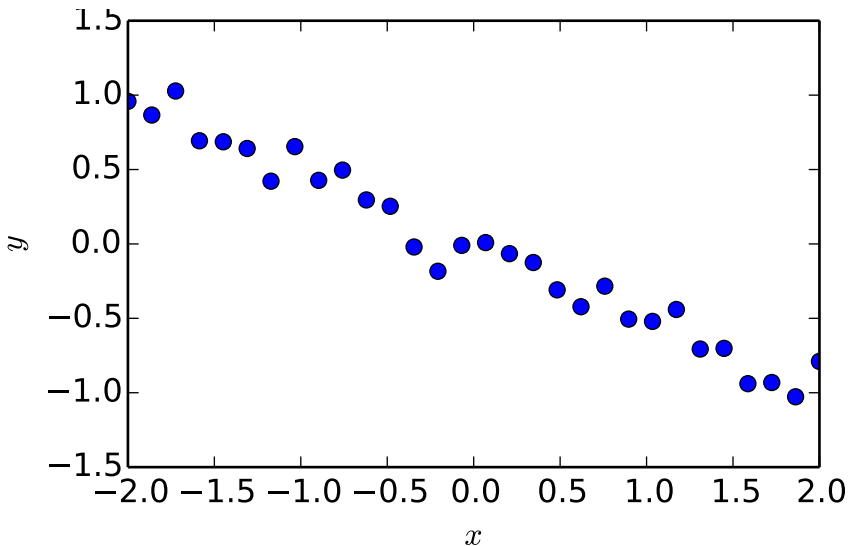
- wenn er nahe bei 1 liegt, dann wachsen  $x$  und  $y$  gemeinsam
- wenn er nahe bei  $-1$  liegt, dann fällt  $y$ , wenn  $x$  wächst
- wenn er nahe bei 0 liegt, dann gibt es keine gemeinsame Tendenz

Datenpaare  
oooooooo●ooo

Lineare Regression  
oooooooooooooooo

Regression im exponentiellen Modell  
oooooooooooo

## Beispiel: Sehr gute Korrelation



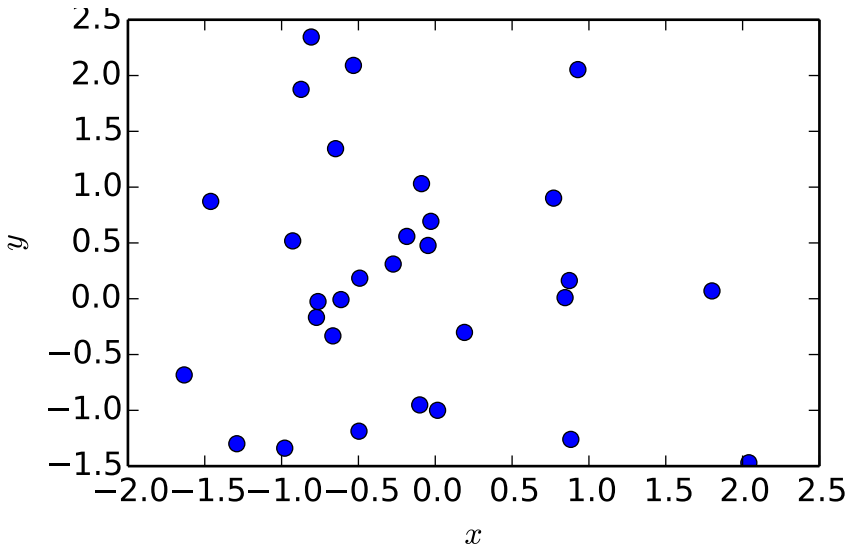
Der empirische Korrelationskoeffizient beträgt  $-0.98$ .

Datenpaare  
oooooooo●oo

Lineare Regression  
ooooooooooooo

Regression im exponentiellen Modell  
ooooooooooooo

## Beispiel: Keine Korrelation

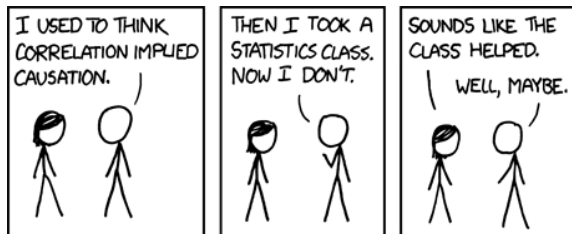


Der empirische Korrelationskoeffizient beträgt  $-0.10$ .

# Korrelation $\neq$ Kausalität

- Wenn der Korrelationskoeffizient von  $x$  und  $y$  nahe 0 liegt, dann gibt es keinen (linearen) kausalen Zusammenhang zwischen ihnen
- Man kann aber im umgekehrten Fall von einem Korrelationskoeffizienten nahe bei 1 nicht auf einen kausalen Zusammenhang schließen
- Zum Beispiel nimmt seit Jahrzehnten in Deutschland sowohl die Zahl der Geburten als auch die Zahl der Störche ab
- Der kausale Zusammenhang ist aber umstritten

xkcd



Quelle: <http://xkcd.com/552>

# Korrelation und lineare Regression

Wenn eine Korrelation zwischen zwei Datensätzen besteht, dann entstehen quantitative Fragen

- Wie sehr steigt oder fällt  $y$  in Abhängigkeit von  $x$ ?
- Welchen Wert  $y$  erwartet man für gegebenes  $x$ ?

# Lineare Regression

- “Lineare Regression”: Bestimmung einer Regressionsgeraden
- “linear”: auf einer Gerade liegend
- “Gerade”: Funktionsvorschrift

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

Hierbei ist

- $m$  die Steigung der Geraden
- $b$  der Ordinatenabschnitt der Geraden
- Die Regressionsgerade ist die Gerade mit der bestmöglichen Annäherung an die Datenpunkte
- “bestmöglich” bedeutet

$$\sum_{j=1}^n (m \cdot x_j + b - y_j)^2 \stackrel{!}{=} \min$$

# Formel für die lineare Regression

- Gegeben: Datenpaare  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Gesucht: Regressionsgerade  $y = m \cdot x + b$
- Rechenvorschrift:

$$m = \frac{\text{covar}_{\text{emp}}(x, y)}{s_x^2}$$

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

- Dabei:
  - $\bar{x}$  und  $\bar{y}$ : arithmetisches Mittel von  $x$  und  $y$
  - $s_x^2$  Varianz von  $x$
  - $\text{covar}_{\text{emp}}(x, y)$  empirische Kovarianz von  $x$  und  $y$

## Beispiel: Blutdruckdaten

$x$  ist das Alter,  $y$  der Blutdruck der Probanden. Hier jetzt die Daten für die gesamte Stichprobe

$$\text{covar}_{\text{emp}} = 349 \quad s_x = 15.2$$

Steigung

$$m = \frac{349}{15.2^2} = 1.51$$

Ordinatenabschnitt

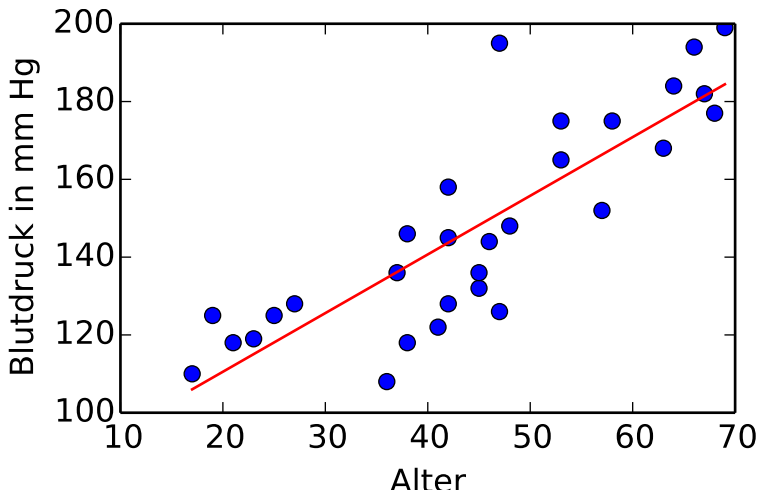
$$b = 148 - 1.51 \cdot 44.8 = 80.4$$

Interpretation:

- Pro Jahr steigt der Blutdruck um 1.5 *mm* Hg
- $b$  hat hier keine Bedeutung, denn nahe bei  $x = 0$  wurden keine Daten erhoben
- Bei einem 50jährigen erwartet man einen Blutdruck von

$$1.51 \cdot 50 + 80.4 = 155.9$$

## Beispiel: lineare Regression der Blutdruckdaten

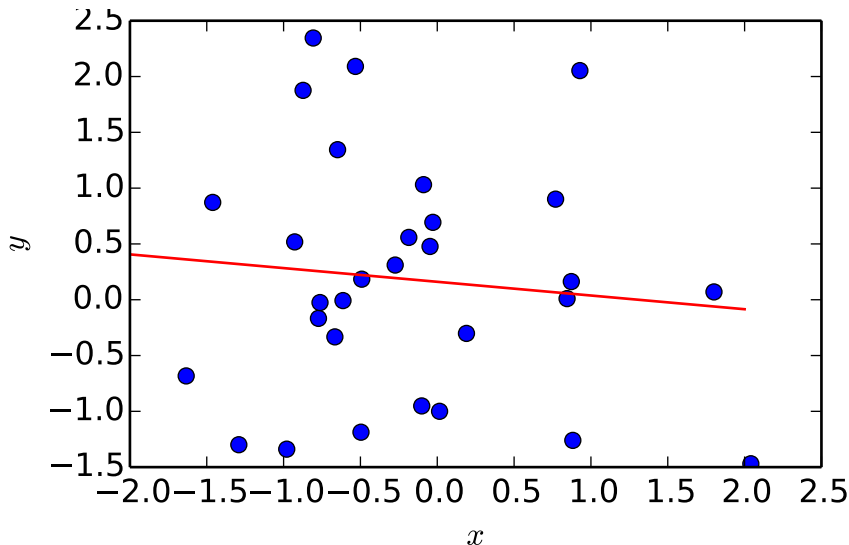


Die Regressionsgerade ist  $y = 1.51 \cdot x + 80.4$  (in *mm Hg*)

# Zu beachten

- Die *Extrapolation* über den Bereich des ursprünglichen Datensatzes hinaus ist unzulässig
- Die Steigung der Regressionsgerade ist nicht gleich der Steigung der Geraden durch die beiden äußersten Punkte
- Der Algorithmus bestimmt immer eine Regressionsgerade
- auch wenn sie keinen Sinn macht

## Beispiel: Schlechte Korrelation mit Regressionsgerade



Die Regressionsgerade ist bedeutungslos

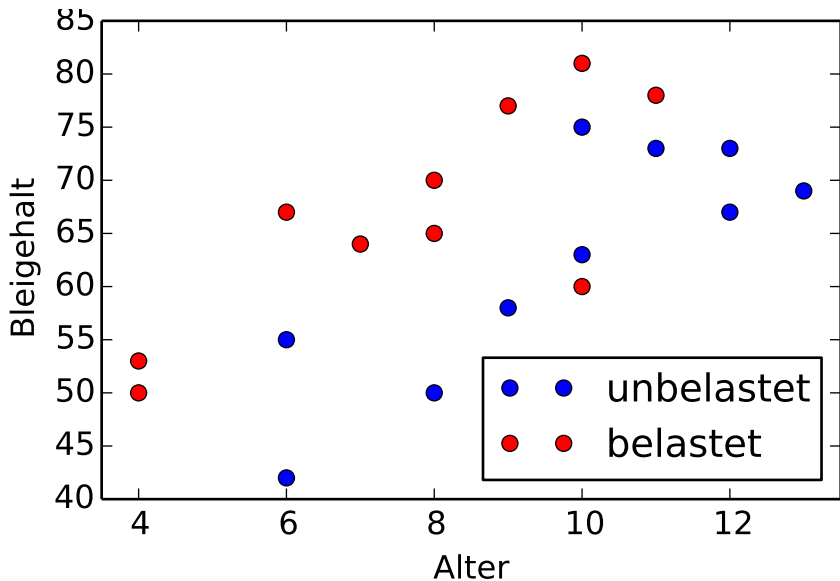
# Durchgerechnetes Beispiel: Bleibelastung im Gewebe von Ratten

- kontaminiertes Gelände: fange 10 Ratten
- unbelastetes Vergleichsgelände: fange 10 Ratten
- für jede Ratte wurden ihr Alter in Monaten und der Bleigehalt im Gewebe bestimmt

## Beispiel: Bleibelastung im Gewebe von Ratten

Vergleichsgelände		kontaminiertes Gelände	
Alter	Belastung	Alter	Belastung
10	63	8	70
12	67	10	60
6	55	4	53
6	42	4	50
11	73	9	77
11	69	11	78
12	73	10	81
10	75	8	65
9	58	7	64
8	50	6	67

## Plot zum Beispiel "Bleibelastung"



# Mittelwerte und Korrelationskoeffizienten

- Tiere von Vergleichsgelände
  - Mittleres Alter: 9.7 Monate
  - Mittelere Bleibelastung: 62.5 Einheiten
  - Korrelationskoeffizient: 0.83
- Tiere von kontaminiertem Gelände
  - Mittleres Alter: 7.7 Monate
  - Mittelere Bleibelastung: 66.5 Einheiten
  - Korrelationskoeffizient: 0.80
- Es gibt einen Unterschied in der Bleibelastung; er ist aber nicht sehr überzeugend
- Bleibelastung steigt mit dem Alter
- Beobachtung: Die Tiere, die in dem belasteten Gelände gefangen wurden, sind im Schnitt jünger

# Lineare Regression für die Bleidaten

- Vergleichsgelände:

$$m_1 = 3.72, \quad b_1 = 26.4$$

- Kontaminiertes Gelände:

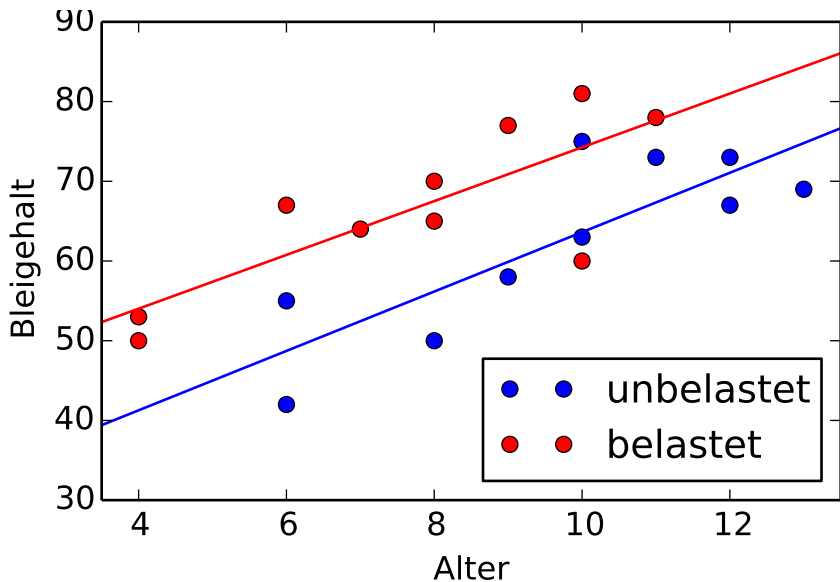
$$m_2 = 3.37, \quad b_2 = 40.5$$

- Die beiden Steigungen sind ähnlich
- Das mittlere Alter über beide Gruppen ist 8.7 Monate.
- Vergleiche die Werte der Regression für dieses Alter

$$m_1 \cdot 8.7 + b_1 = 58.8 \quad m_2 \cdot 8.7 + b_2 = 69.9$$

- Dieser Unterschied ist sehr viel überzeugender

## Zweite Grafik zum Beispiel "Bleibelastung"



# Gendatenbank

Anzahl der Sequenzen in der Gendatenbank des NCBI  
(National Center for Biotechnology Information)

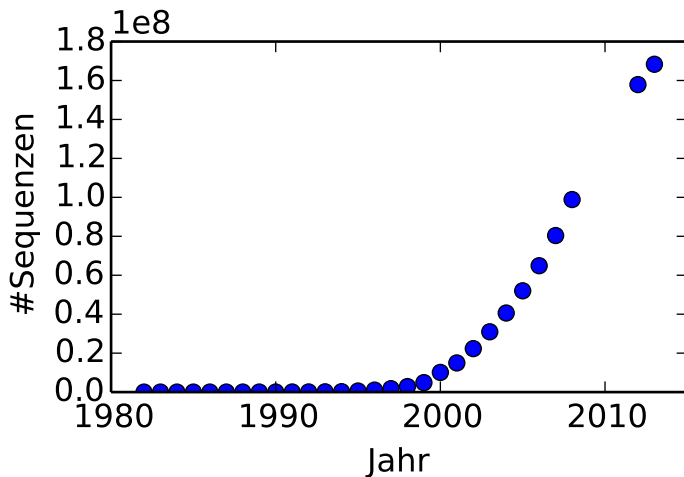
Jahr	Sequenzen	Jahr	Sequenzen	Jahr	Sequenzen
1982	606	1992	78 608	2002	22 318 883
1983	2 427	1993	143 492	2003	30 968 418
1984	4 175	1994	215 273	2004	40 604 319
1985	5 700	1995	555 694	2005	52 016 762
1986	9 978	1996	1 021 211	2006	64 893 747
1987	14 584	1997	1 765 847	2007	80 388 382
1988	20 579	1998	2 837 897	2008	98 868 465
1989	28 791	1999	4 864 570	2012	157 889 737
1990	39 533	2000	10 106 023	2013	168 335 396
1991	55 627	2001	14 976 310		

Datenpaare  
oooooooooooo

Lineare Regression  
oooooooooooo

Regression im exponentiellen Modell  
●ooooooooo

# Anzahl der Sequenzen in der Gendatenbank



# Halblogarithmische Darstellung

Bei halblogarithmischer Darstellung

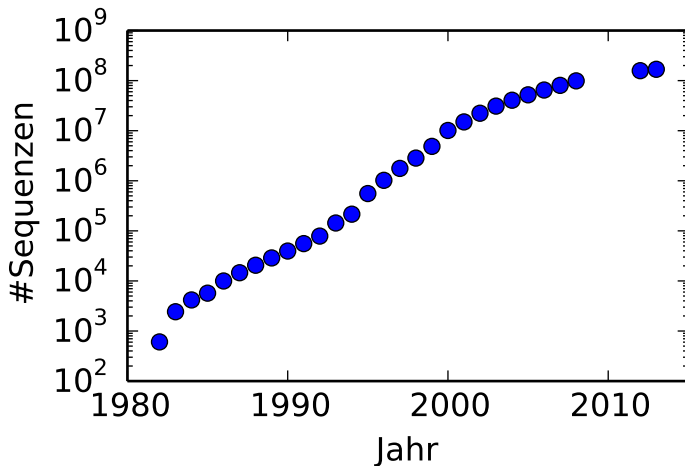
- ist die  $x$ -Achse linear skaliert: Gleiche absolute Zuwächse pro Längeneinheit
- ist die  $y$ -Achse logarithmisch skaliert: Gleiche relative Zuwächse pro Längeneinheit
- Das bedeutet: Der Logarithmus der Daten wird angezeigt, und die  $y$ -Achse wird entsprechend unterteilt
- Exponentiell wachsende Daten liegen bei halblogarithmischer Darstellung annähernd auf einer wachsenden Geraden, exponentiell fallende auf einer fallenden

Datenpaare  
oooooooooooo

Lineare Regression  
oooooooooooo

Regression im exponentiellen Modell  
ooo●oooooooo

# Halblogarithmischer Graph der Anzahl der Sequenzen



# Regression im exponentiellen Modell

- Lineares Modell: in gleichen Zeitabständen gleiche absolute Zuwächse
- Exponentielles Modell: in gleichen Zeitabständen gleiche relative Zuwächse
- Biologische Wachstums- oder Abklingprozesse verlaufen meistens exponentiell
- Aufgabe der Regression im exponentiellen Modell ist es, bei Wachstumsprozessen die Verdoppelungszeit und bei Abklingprozessen die Halbwertszeit zu bestimmen
- Dies geschieht, indem man die Werte logarithmiert und dann deren lineare Regression berechnet

# Regression im exponentiellen Modell

- $x$  die Zeit,  $z$  Daten, die exponentiell wachsen (bzw. abklingen)
- Modellgleichung für Wachstumsprozess:

$$z = c \cdot e^{m \cdot x}$$

- logarithmierte Modellgleichung

$$y = \ln(z) = \ln(c) + m \cdot x$$

- bestimme diese Gerade durch lineare Regression
- wenn  $m < 0$ , dann Abklingprozess

# Halbwerts- bzw. Verdoppelungszeit

- Modell eines Wachstumsprozesses

$$z = c \cdot e^{m \cdot x}$$

- Verdoppelungszeit  $t$  bestimmt durch

$$e^{m \cdot t} = 2$$

- Also

$$t = \frac{\ln 2}{m}$$

- Bei Abklingprozessen ist  $m < 0$ , dann ist

$$t = -\frac{\ln 2}{m}$$

die Halbwertszeit

# Gendatenbank: Regression im exponentiellen Modell

$z$  ist die Anzahl der Sequenzen in der Gendatenbank

Jahr	$\ln(z)$	Jahr	$\ln(z)$	Jahr	$\ln(z)$
1982	6.41	1992	11.27	2002	16.92
1983	7.79	1993	11.87	2003	17.25
1984	8.34	1994	12.28	2004	17.52
1985	8.65	1995	13.23	2005	17.77
1986	9.21	1996	13.84	2006	17.99
1987	9.59	1997	14.38	2007	18.20
1988	9.93	1998	14.86	2008	18.41
1989	10.27	1999	15.40	2012	18.88
1990	10.58	2000	16.13	2013	18.94
1991	10.93	2001	16.52		

# Gendatenbank: Fortsetzung der Regression

- Lineare Regression für  $y = \ln(z)$

$$\bar{x} = 1996.21$$

$$\bar{y} = 13.564$$

$$s_x^2 = 78.884$$

$$\text{covar}_{\text{emp}}(x, y) = 33.960$$

- Also

$$m = \frac{\text{covar}_{\text{emp}}(x, y)}{s_x^2} = 0.4305$$

- Dazu gehört die Verdopplungszeit

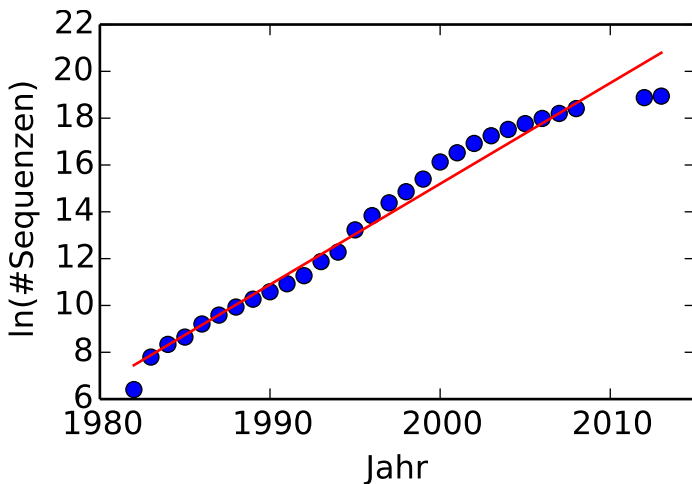
$$x_d = \frac{\ln(2)}{m} = 1.61 \text{Jahre}$$

Datenpaare  
oooooooooooo

Lineare Regression  
oooooooooooo

Regression im exponentiellen Modell  
ooooooooo●o

# Lineare Regression der logarithmierten Daten



Datenpaare  
oooooooooooo

Lineare Regression  
oooooooooooo

Regression im exponentiellen Modell  
oooooooooooo●

# Regression im exponentiellen Modell, halblogarithmische Darstellung

