Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

21. Oktober 2010

Datenpaare

- Häufig erhebt man zwei Merkmale, um deren Abhängigkeit zu erforschen
- beispielsweise könnte man Alter und Blutdruck messen
- mathematisch stellt man solch ein Ergebnis als Menge von Datenpaaren dar

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

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)

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.
- Diesen berechnet man wiederum aus der empirischen Kovarianz.
- Wenn Korrelation besteht, kann sie durch die Methode der linearen Regression quantifiziert werden.

Empirische Kovarianz

- Wir haben n Datenpaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- das arithmetische Mittel der x_j ist $\overline{x}$, das der y_j ist $\overline{y}$
- die *empirische Kovarianz* von x und y ist

$$\begin{aligned}\operatorname{covar}_{\text{emp}}(x, y) \\ &= \frac{1}{n-1}\big((x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots \\ &\qquad + (x_n - \overline{x})(y_n - \overline{y})\big)\end{aligned}$$

- Formel ohne Pünktchen

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

Empirischer Korrelationskoeffizient

- s_x sei die Stichprobenstreuung der x_j und s_y die Stichprobenstreuung der y_j
- dann ist der *empirische Korrelationskoeffizient* gleich

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

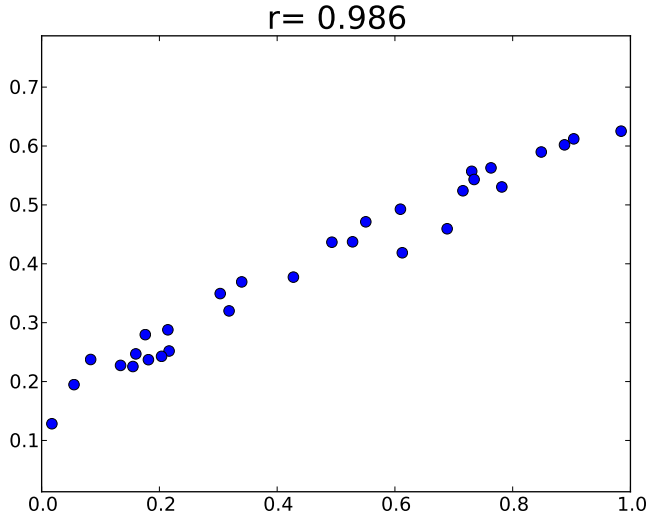
- ausgeschrieben bedeutet das

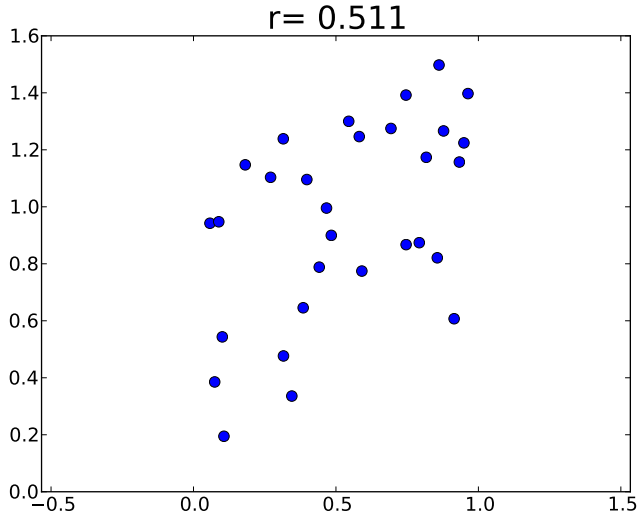
$$r = \frac{\sum_{j=1}^{n} (x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\left(\sum_{j=1}^{n} (x_j - \overline{x})^2\right)\left(\sum_{j=1}^{n} (y_j - \overline{y})^2\right)}}$$

Randfälle

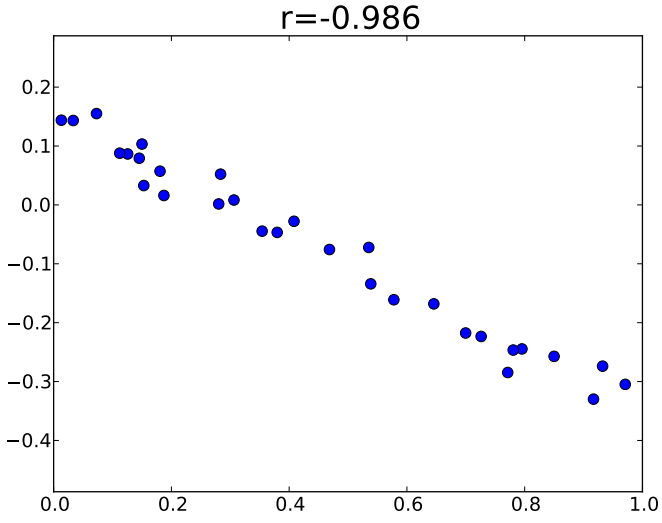
- wenn $y = x$, dann ist die Kovarianz von x und y (also die Kovarianz von x mit sich selbst) gleich der Varianz von x
- also ist in diesem Fall der Korrelationskoeffizient gleich 1
- wenn $y = -x$, dann ist die Kovarianz von x und y (also die Kovarianz von x mit $-x$) gleich dem Negativen der Varianz von x
- in diesem Fall ist der Korrelationskoeffizient gleich -1
- wenn alle x_j oder alle y_j gleich sind, dann sind Kovarianz und Korrelationskoeffizient gleich 0
- der Korrelationskoeffizient liegt immer zwischen -1 und 1

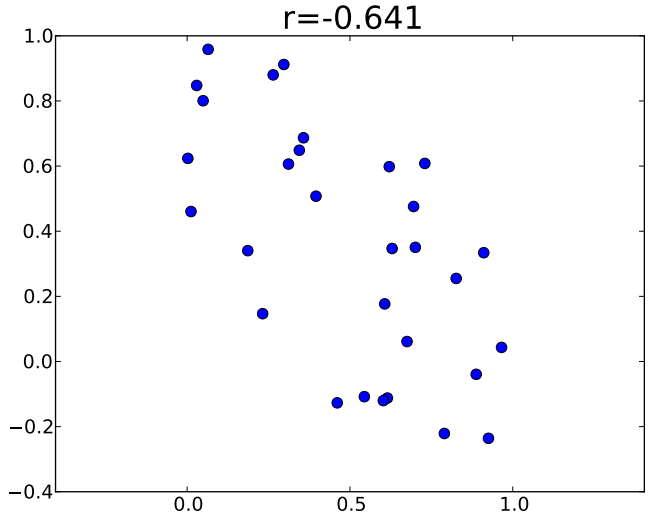
positiver Korrelationskoeffizient



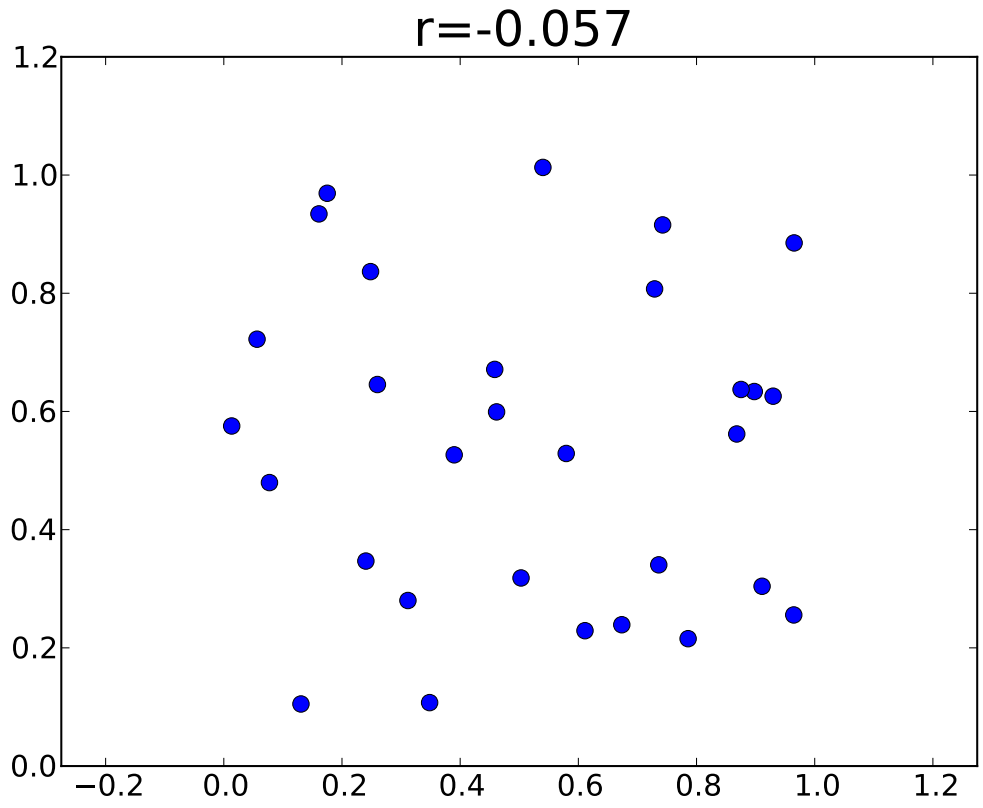


negativer Korrelationskoeffizient





Korrelationskoeffizient nahe Null



Interpretation

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

- 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

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

Lineare Regression

- "linear": auf einer Gerade liegend
- "Gerade": Funktionsvorschrift

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

Hierbei ist

- m die Steigung der Geraden
- b der Ordinatenabschnitt der Geraden

- "lineare Regression": Annäherung der Daten durch die bestmögliche Gerade

Formel für die lineare Regression

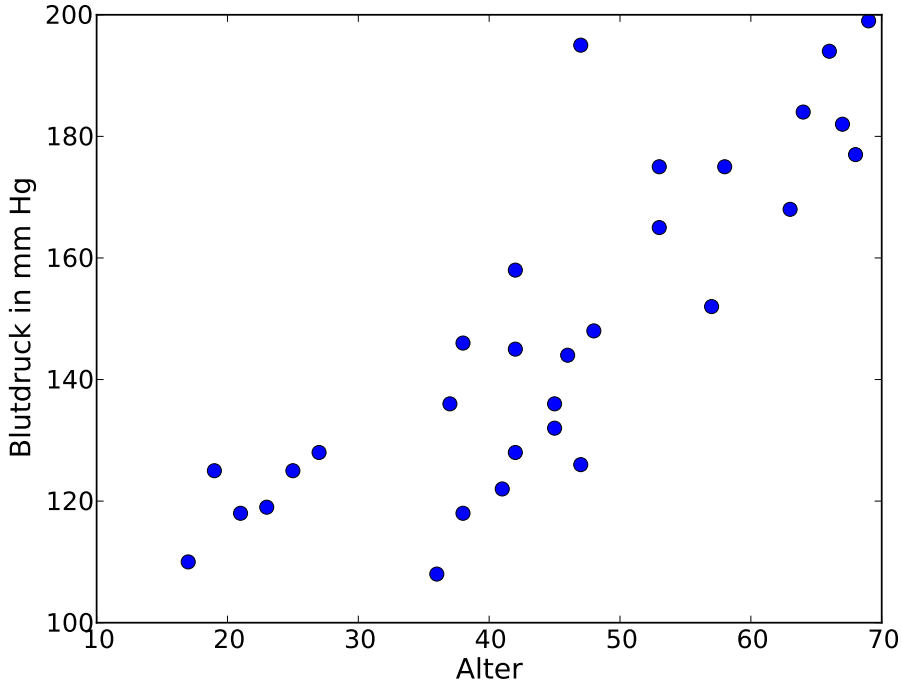
- Gegeben: Datenpaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Gesucht: Regressionsgerade $y = m \cdot x + b$
- Rechenvorschrift:

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

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

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

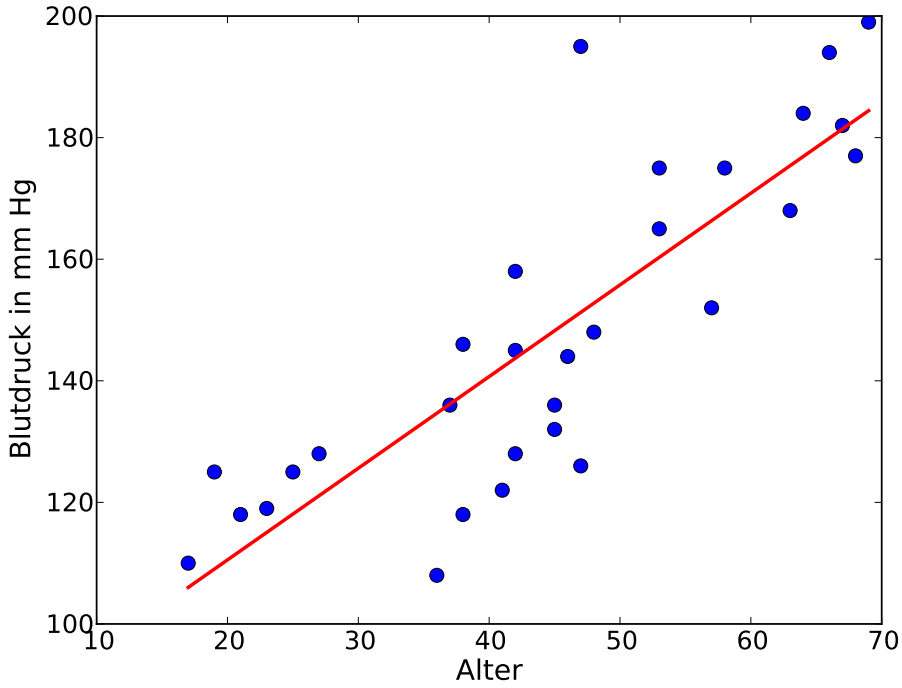
Beispiel: Blutdruckdaten



Blutdruckwerte von 30 gesunden Männern. Der empirische Korrelationskoeffizienten beträgt

$$r = 0.84$$

Beispiel: lineare Regression der Blutdruckdaten



Die Regressionsgerade ist

$$y = 1.5 \cdot x + 80$$

Einheit: mm Hg

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. schrumpfen)
- Modellgleichung für Wachstumsprozess:

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

  Modellgleichung für Abklingprozess:

$$z = c \cdot e^{-A \cdot x}$$

- logarithmierte Modellgleichung

$$y = \ln(z) = \ln(c) \pm A \cdot x$$

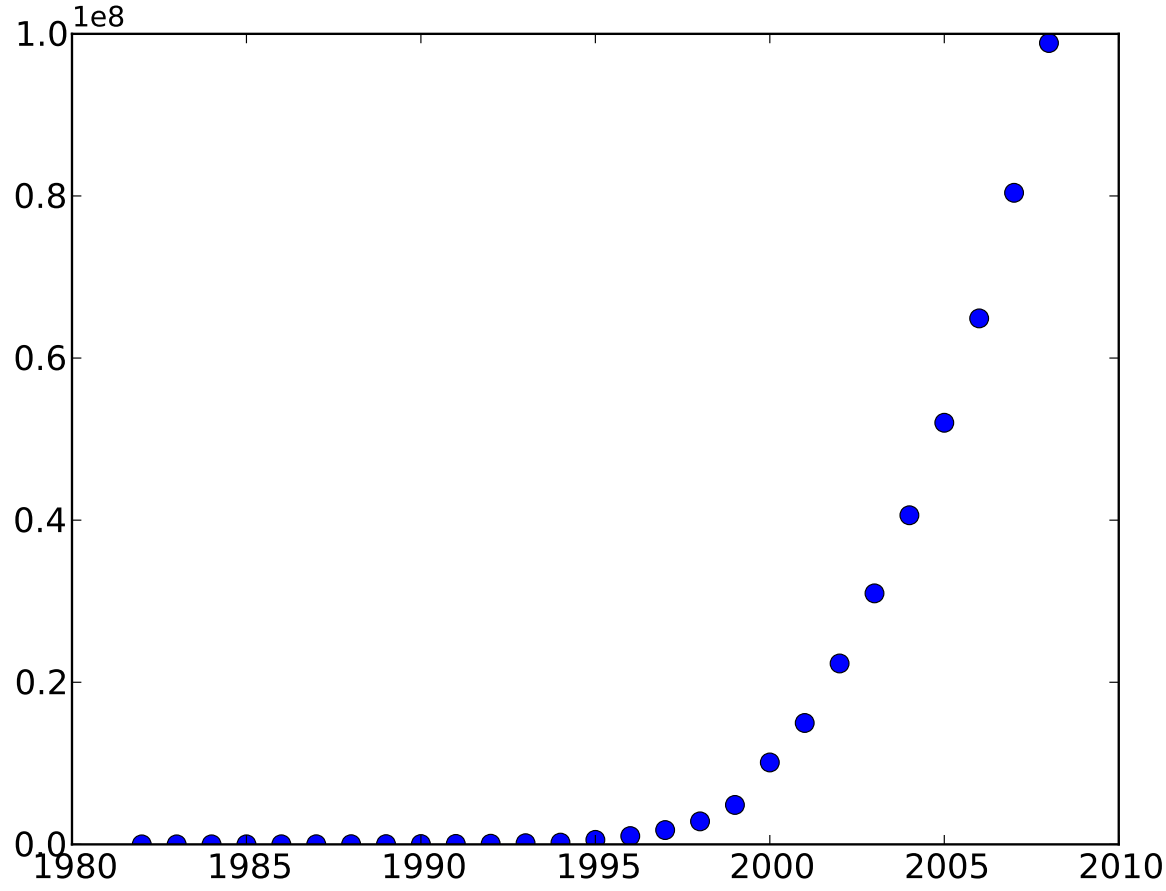
- bestimme diese Gerade durch lineare Regression

Gendatenbank

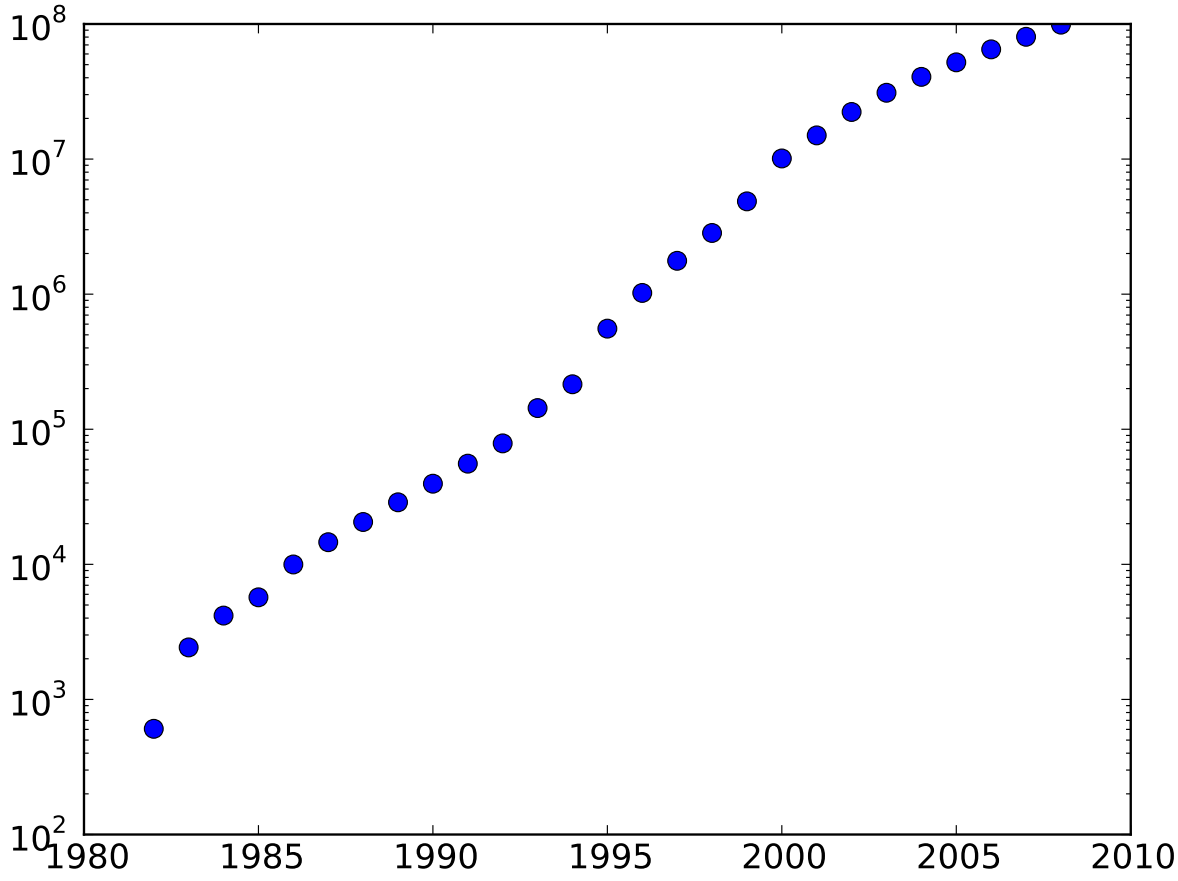
Anzahl der Sequenzen in der Gendatenbank des NCBI

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

Anzahl der Sequenzen in der Gendatenbank



Halblogarithmischer Graph der Anzahl der Sequenzen



Der halblogarithmische Graph von z zeigt $\ln(z)$, aber mit der Skaleneinteilung von z.

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

Gendatenbank: Fortsetzung der Regression

- Lineare Regression für $y = \ln(z)$

$$\overline{x} = 1995 \qquad \overline{y} = 13.17$$

$$s^2 = 63 \quad \text{covar}_{\text{emp}}(x, y) = 29.38$$

- Also

$$m = \frac{\text{covar}_{\text{emp}}(x, y)}{s_x^2} = 0.4662$$

- Dazu gehört die Verdopplungszeit

$$x_d = \frac{\ln(2)}{m} = 1.487$$

Lineare Regression der logarithmierten Daten