3. Beschreibung von mehrdimensionalen Daten

3.1 Grafische Darstellungen einer Stichprobe von Zahlenpaaren

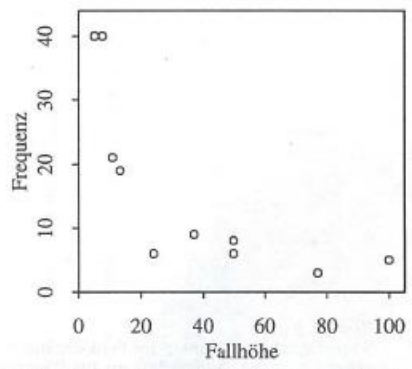
Werden bei einer Stichprobe zwei verschiedene Merkmale x_i und y_i beobachtet, so spricht man von einer bivariaten Stichprobe. Dabei stellt sich die Frage, ob es einen Zusammenhang, eine Abhängigkeit zwischen den beiden Grössen gibt. In diesem Fall sagt man, dass die beiden Variablen korreliert sind.

Beispiele:
Druck und Volumen bei einem physikalischen Experiment, Spraydosen-Treibgasverbrauch und mittlerer Ozongehalt der Atmosphäre in einem Jahr, Grösse des Vaters und des Sohnes bei einem Vater-Sohn-Paar.

Die Daten (x_i, y_i) können als Punkte in der Koordinatenebene dargestellt werden. Diese Darstellung heisst Streudiagramm (englisch: Scatterplot). Sie erleichtert es, bei Zahlenpaaren Trends, Muster, Ausreisser zu erkennen.

Beispiel:
Das Rauschen von Wasserfällen (Quelle: Science 164,1969, p. 1513-1514)

Bei verschiedenen Wasserfällen wurde die Höhe und die dominierende Frequenz im Spektrum der Bodenvibrationen gemessen. Im Streudiagramm ist ersichtlich, dass hohe Frequenzen mit niedrigen Höhen einhergehen und umgekehrt.



Name	x_i Höhe (m)	y_i Frequenz (Hz)
Lower Yellowstone	100	5
Yosemite	77	3
Canadian Niagara	50	6
American Niagara	50	8
Upper Yellowstone	37	9
Gullfoss (Lower)	24	6
Firehole	13.3	19
Godafoss	10.9	21
Gullfoss (Upper)	7.7	40
Fort Greeley	5.2	40

Bild 3.1.d Streudiagramm für Fallhöhe und Schwingungsfrequenz bei Wasserfällen

Sind mehr als zwei Variablen von Interesse, dann wird - wie im folgenden Beispiel - in der so genannten Scatterplot-Matrix jeweils für zwei Merkmale ein Streudiagramm abgebildet.

Beispiel:
Es sind die Streudiagramme der mittleren Monatsrenditen der vier Aktien: Volkswagen; BASF, Siemens, Münchner Rück und zusätzlich der Libor-Zins abgebildet. Der Libor-Zins (**L**ondon **I**nter**b**ank **O**ffered **r**ate) ist der Zinssatz, zu dem sich Geschäftsbanken gegenseitig Geld mit einer Laufzeit bis zu einem Jahr leihen. In der Matrix ist zu erkennen, dass die Höhe der Renditen für Aktien untereinander deutlich zusammenhängen, während kein deutlicher Zusammenhang zwischen Aktienrendite und Zins erkennbar ist.

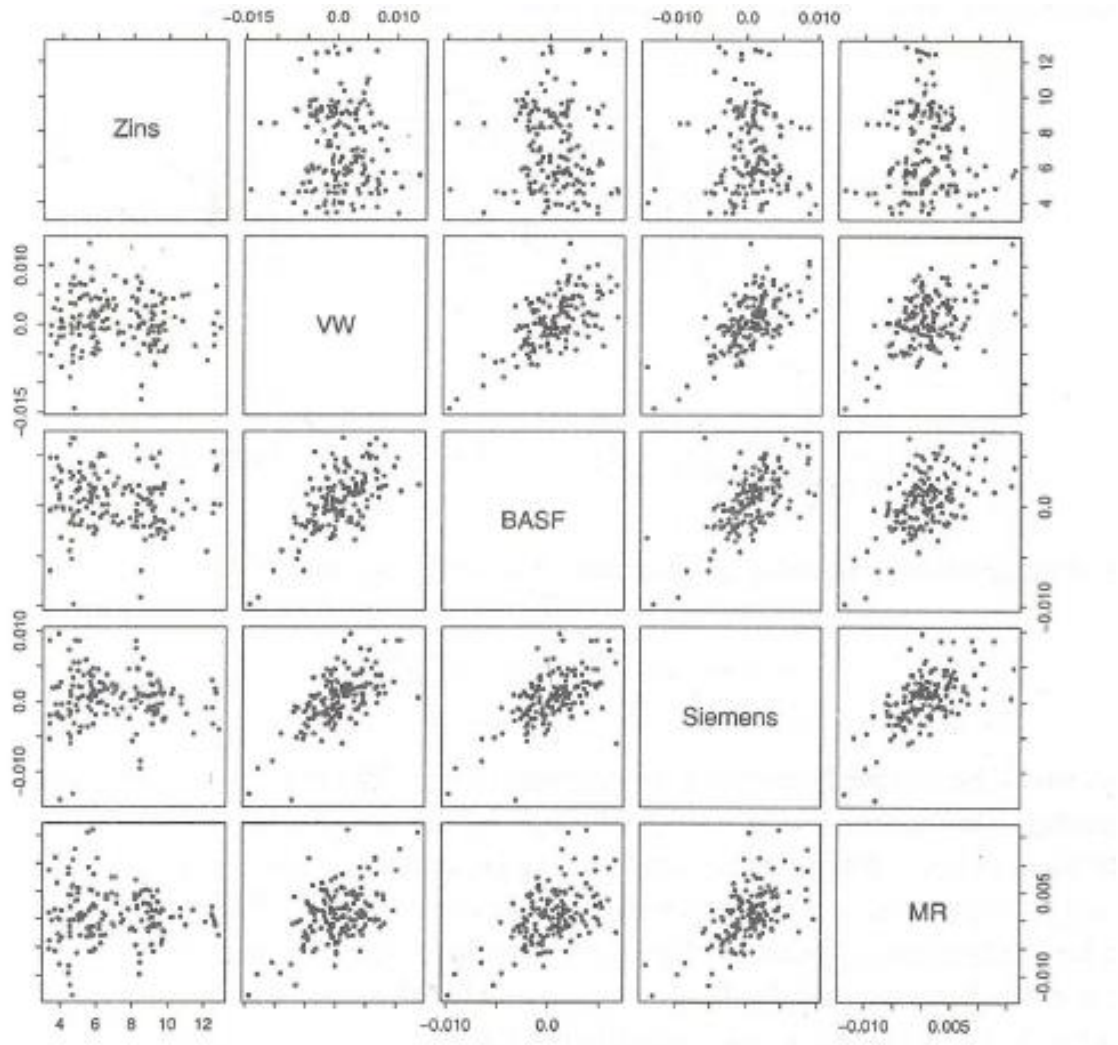


ABBILDUNG 3.10: Streudiagramm-Matrix der mittleren Monatsrenditen für Aktien und des Zinses in Form des Libors

Quelle: Fahrmeier: Statistik, Springer, 2001

3.2 Kennzahlen für bivariate Stichproben

Zunächst können die bekannten Kennzahlen für univariate Stichproben bestimmt werden:

Empirischer Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad 1)$$

Varianz s_x^2 bzw. Standardabweichung s_x

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \qquad 2)$$

bzw.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \qquad 2´)$$

Als Gesamtmass wird die folgende sogenannte Produkt-Momentenkorrelation r_{xy} (englisch: Pearson-Correlation) verwendet:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_I - \bar{y})^2}} \qquad 5)$$

Definiert man als Empirische Kovarianz s_{xy}

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) \;= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right) \qquad 6)$$

dann kann die Produkt-Momentenkorrelation r_{xy} mit 6) kurz folgendermassen geschrieben werden:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5')$$

Schreibt man die Definition 5) in der folgenden Form um, so ist zu erkennen, dass die Korrelation nicht vom Nullpunkt der Skalen von x bzw. y und von den entsprechenden Masseinheiten abhängt:

$$r_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot \frac{(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_I - \bar{y})^2}}$$

Führt man analog zum Vorgehen bei der Normalverteilung die folgenden neuen standardisierten Variablen ein

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

bzw.

$$v_i = \frac{y_i - \bar{y}}{s_y}$$

und bildet man mit diesen Werten die Vektoren $\vec{u}$ und $\vec{v}$, so vereinfacht sich wegen $|\vec{u}| = |\vec{v}| = \sqrt{n-1}$ die Formel und die Summe kann als Skalarprodukt aufgefasst werden.

$$r_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} u_i v_i = \frac{1}{n-1} \cdot (\vec{u} \cdot \vec{v}) \qquad 7)$$

Da die Definition der Korrelation nicht von der Wahl der linearen Skala abhängt, gilt zudem: $r_{xy} = r_{uv}$

Wenn beide Koordinaten das gleiche Vorzeichen haben (die zugehörigen Punkte liegen dann im verschobenen Koordinatensystem mit Zentrum im Schwerpunkt S $(\bar{x}, \bar{y})$ im ersten und dritten Quadranten), so deutet dies auf einen „positiven“ Zusammenhang zwischen den beiden Variablen, und entsprechend bedeuten Koordinaten mit ungleichem Vorzeichen auf einen „negativen“ Zusammenhang hin (die zugehörigen Punkte liegen im zweiten und vierten Quadranten).

Die aufgeführten Umformungen bei 1) bis 7) ermöglichen die Berechnung aller Kennzahlen aus geeigneten Summen. Allerdings besteht bei dieser Berechnungsart die Gefahr von Rundungsfehlern.

Beispiel Wasserfälle:

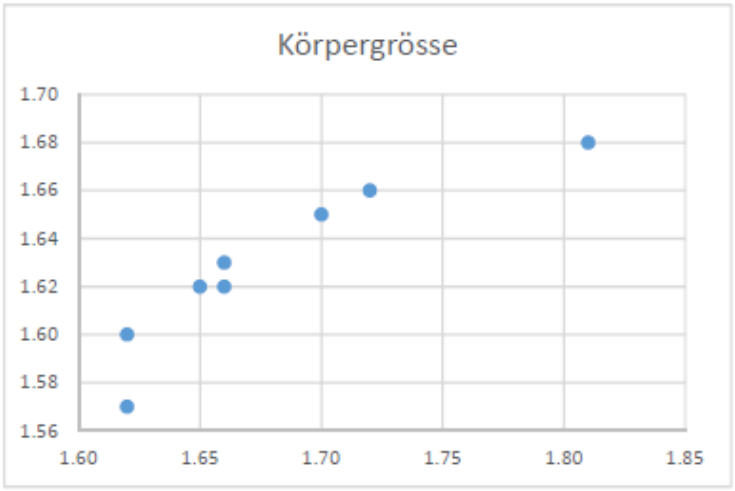
Hoehe [m]	Frequenz [Hz]							
x_i	y_i	x_i^2	y_i^2	$x_i y_i$		u_i	v_i	$u_i v_i$
100	5	10000.00	25	500.0		1.9559956	-0.75911772	-1.48483092
77	3	5929.00	9	231.0		1.23607403	-0.90100889	-1.11371369
50	6	2500.00	36	300.0		0.39094871	-0.68817214	-0.26904001
50	8	2500.00	64	400.0		0.39094871	-0.54628098	-0.21356784
37	9	1369.00	81	333.0		-0.01596348	-0.4753354	0.00758801
24	6	576.00	36	144.0		-0.42287567	-0.68817214	0.29101125
13.3	19	176.89	361	252.7		-0.7577957	0.23412042	-0.17741545
10.9	21	118.81	441	228.9		-0.83291795	0.37601158	-0.3131868
7.7	40	59.29	1600	308.0		-0.93308095	1.72397763	-1.60861069
5.2	40	27.04	1600	208.0		-1.0113333	1.72397763	-1.74351598
375.1	157	23256.03	4253	2905.6	Summe	0	0	-6.62528213
37.51					$\bar{x}$			
	15.7				$\bar{y}$			
		1020.67						
		31.95			s_x			
			198.68					
			14.10		s_y			
				-331.50	s_{xy}			
				-0.736	r_{xy}		r_{xy}	-0.736

Ein weiteres Beispiel:
In der Klasse 1C der Kantonsschule Zofingen wurden am 21.3.2003 die Körpergrössen der Paare Tochter- Mutter bzw. Sohn-Vater gemessen.

Körpergrösse 1C 21.03.2003

Tochter x	**Mutter y**
1.62	1.57
1.62	1.60
1.65	1.62
1.66	1.62
1.66	1.63
1.70	1.65
1.72	1.66
1.81	1.68

Korrelation **0.911**

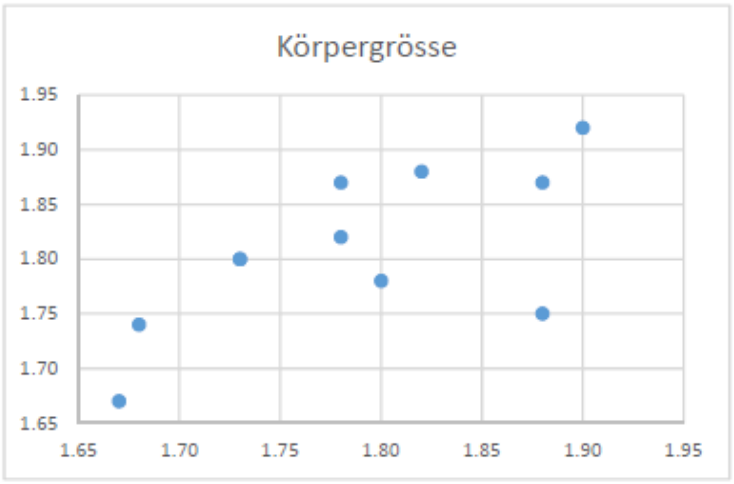


Teilresultate

$\bar{\bar{x}}$	1.680
sd_x	0.063
$\bar{\bar{y}}$	1.629
sd_y	0.035
Kovarianz	0.002
Korrelation	**0.911**

Sohn x **Vater y** **Korrelation** **0.669**

1.67	1.67
1.68	1.74
1.73	1.80
1.73	1.80
1.78	1.87
1.78	1.82
1.80	1.78
1.82	1.88
1.88	1.75
1.88	1.87
1.90	1.92



Teilresultate

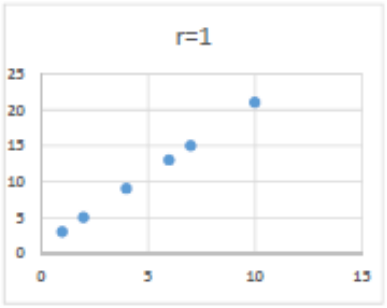
$\bar{\bar{x}}$	1.786
sd_x	0.079
$\bar{\bar{y}}$	1.809
sd_y	0.073
Kovarianz	0.004
Korrelation	**0.669**

Hypothese:
Die Körpergrössen des Paars Vater-Sohn korrelieren schwächer als beim Paar Mutter Tochter.
Im Kapitel „Schliessende Statistik“ wird die Frage behandelt, wie eine solche Hypothese überprüft werden kann.

In den folgenden Abbildungen sind die Scatterplots von acht Stichproben mit verschiedenen Korrelationskoeffizienten dargestellt. Zusätzlich sind die Werte der Standardabweichungen von x bzw. y und die Kovarianz angegeben, aus denen sich die Korrelation gemäss 5´) berechnen lässt.

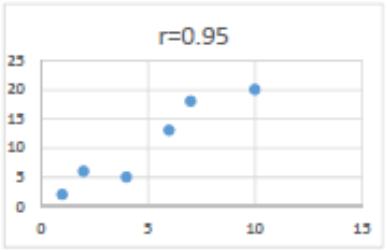
Scatterplot

i	x_i	y_i
1	1	3
2	2	5
3	4	9
4	6	13
5	7	15
6	10	21



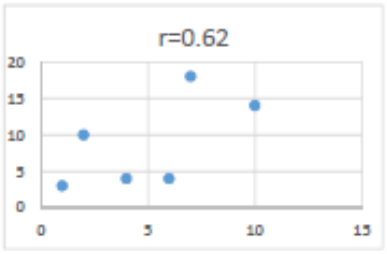
$\bar{x}$	5	$\bar{y}$	11.00
s_x	3.35	s_y	6.69
Kovarianz			22.40
Korrelation			**1.00**

i	x_i	y_i
1	1	2
2	2	6
3	4	5
4	6	13
5	7	18
6	10	20



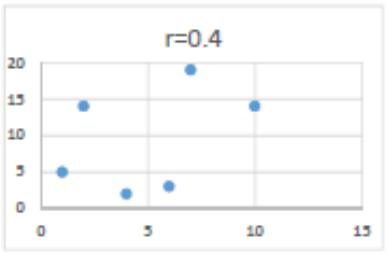
$\bar{x}$	5	$\bar{y}$	10.67
s_x	3.35	s_y	7.42
Kovarianz			23.60
Korrelation			**0.95**

i	x_i	y_i
1	1	3
2	2	10
3	4	4
4	6	4
5	7	18
6	10	14



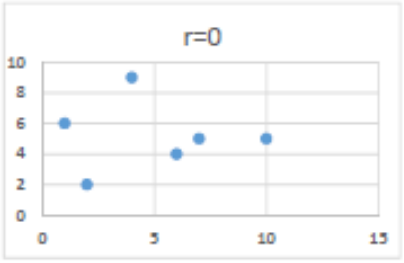
$\bar{x}$	5	$\bar{y}$	8.83
s_x	3.35	s_y	6.21
Kovarianz			12.80
Korrelation			**0.62**

i	x_i	y_i
1	1	5
2	2	14
3	4	2
4	6	3
5	7	19
6	10	14



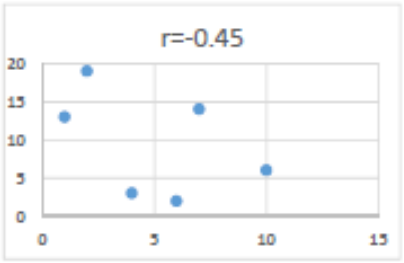
$\bar{x}$	5	$\bar{y}$	9.50
s_x	3.35	s_y	7.06
Kovarianz			9.40
Korrelation			**0.40**

i	x_i	y_i
1	1	6
2	2	2
3	4	9
4	6	4
5	7	5
6	10	5



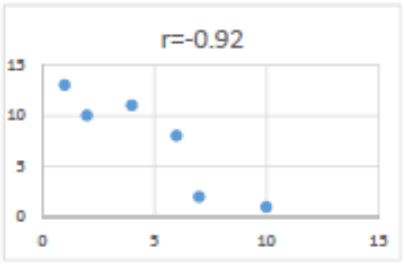
$\bar{x}$	5	$\bar{y}$	5.17
s_x	3.35	s_y	2.32
Kovarianz			0.00
Korrelation			**0.00**

i	x_i	y_i
1	1	13
2	2	19
3	4	3
4	6	2
5	7	14
6	10	6



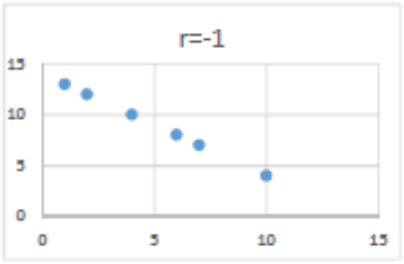
$\bar{x}$	5	$\bar{y}$	9.50
s_x	3.35	s_y	6.83
Kovarianz			-10.40
Korrelation			**-0.45**

i	x_i	y_i
1	1	13
2	2	10
3	4	11
4	6	8
5	7	2
6	10	1



$\bar{x}$	5	$\bar{y}$	7.50
s_x	3.35	s_y	4.93
Kovarianz			-15.20
Korrelation			**-0.92**

i	x_i	y_i
1	1	13
2	2	12
3	4	10
4	6	8
5	7	7
6	10	4



$\bar{x}$	5	$\bar{y}$	9.00
s_x	3.35	s_y	3.35
Kovarianz			-11.20
Korrelation			**-1.00**

Eigenschaften der Produkt-Momenten-Korrelation

a)
Die Korrelation r_{xy} ist +1 für eine Gerade mit positiver Steigung, -1 für eine mit negativer Steigung. Für Werte von r_{xy} nahe bei +1 oder -1 liegen die Punkte eng um eine Gerade. Die Korrelation ist also ein Mass für die Stärke und Richtung eines linearen Zusammenhangs. Ist r_{xy} = 0, so besteht kein linearer Zusammenhang.

b)
Es gilt $|r_{xy}| \leq 1$
Ein einfacher Beweis ergibt sich nach der sogenannten Cauchy-Schwarzschen Ungleichung, die von Ebene und Raum her bereits bekannt ist:

$(\vec{u} \cdot \vec{v})^2 \leq (|\vec{u}||\vec{v}|)^2$

Zum Beweis wählt man für die Berechnung der Korrelation wie in 3.2 die standardisierten Variablen
$u_i = \frac{x_i - \bar{x}}{s_x}$ bzw. $v_i = \frac{y_i - \bar{y}}{s_y}$
und bildet mit diesen Werten die Vektoren $\vec{u}$ und $\vec{v}$. gilt

Wegen der Standardisierung folgt daraus für die Varianz nach 2)
$$s_u^2 = \frac{1}{n-1}|\vec{u}|^2$$
$$s_v^2 = \frac{1}{n-1}|\vec{v}|^2$$

und wegen
$$\left(\sum_{i=1}^{n} u_i\right) = \left(\sum_{i=1}^{n} v_i\right) = 0$$
nach 6) für die Kovarianz
$$s_{uv} = \frac{1}{n-1} \cdot (\vec{u} \cdot \vec{v})$$
und schliesslich für die Produkt-Momenten-Korrelation nach 5´) und Cauchy-Schwarz
$$r_{xy}^2 = r_{uv}^2 = \left(\frac{s_{uv}}{s_u s_v}\right)^2 = \frac{\frac{(\vec{u} \cdot \vec{v})^2}{(n-1)^2}}{\frac{1}{n-1}|\vec{u}|^2 \cdot \frac{1}{n-1}|\vec{v}|^2} \frac{(\vec{u} \cdot \vec{v})^2}{(|\vec{u}| \cdot |\vec{v}|)^2} \leq 1$$

c)
Die folgenden Streudiagramme stellen Stichproben mit gleichem Korrelationskoeffizienten 0.7 dar (Quelle: Chambers, Cleveland, Kleiner and Tukey 1983).
Es ist der grosse Einfluss von allfälligen Ausreissern zu erkennen. In Anwendungen sollte deshalb nicht nur die Korrelation allein, sondern auch das zugehörige Streudiagramm beachtet werden.

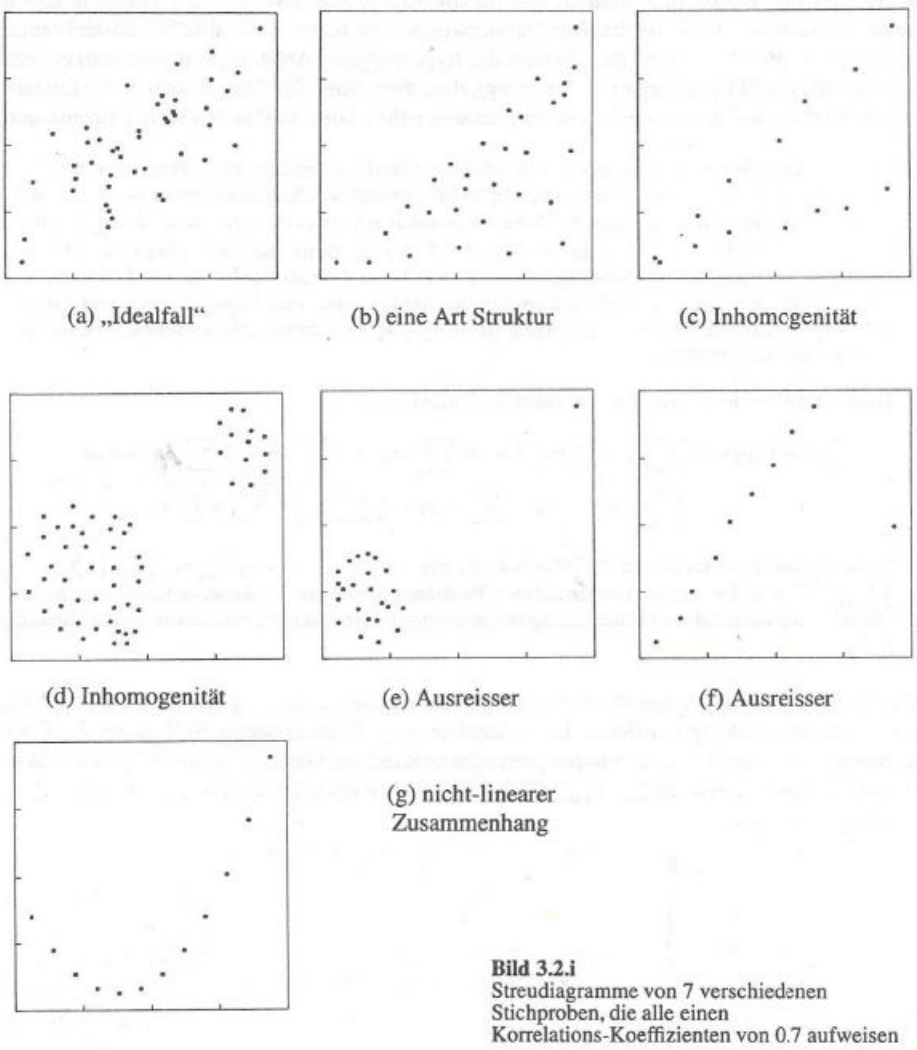


Bild 3.2.i
Streudiagramme von 7 verschiedenen Stichproben, die alle einen Korrelations-Koeffizienten von 0.7 aufweisen

d)

Korrelationen sollten vorsichtig interpretiert werden. Selbst ein offensichtlicher Zusammenhang braucht keinen ursächlichen Zusammenhang zu bedeuten. Es kann nämlich sein, dass beide Grössen von einer dritten Grösse abhängen.

Dass die Anzahl der bei einem Brand eingesetzten Feuerwehrleute hoch korreliert ist mit dem angerichteten Schaden, bedeutet nicht, dass die Feuerwehrleute besser daheim bleiben würden.
Heikel sind insbesondere Zeitreihen mit linearer Komponente. Im bekannten Beispiel ist die hohe Korrelation zwischen der Zahl der Störche und der Zahl der Geburten zwischen 1900 und 1970 kein Beweis, dass der Klapperstorch die Babys bringt.
Es besteht z.B. eine positive Korrelation zwischen Körpergewicht und manueller Geschicklichkeit bei Schulkindern, denn Kinder werden mit steigendem Alter im Mittel schwerer und gleichzeitig geschickter.
Bei Männern besteht zwischen dem Einkommen und der Zahl der Haare auf dem Kopf eine negative Korrelation. Die erklärende Variable ist auch in diesem Fall das Alter.

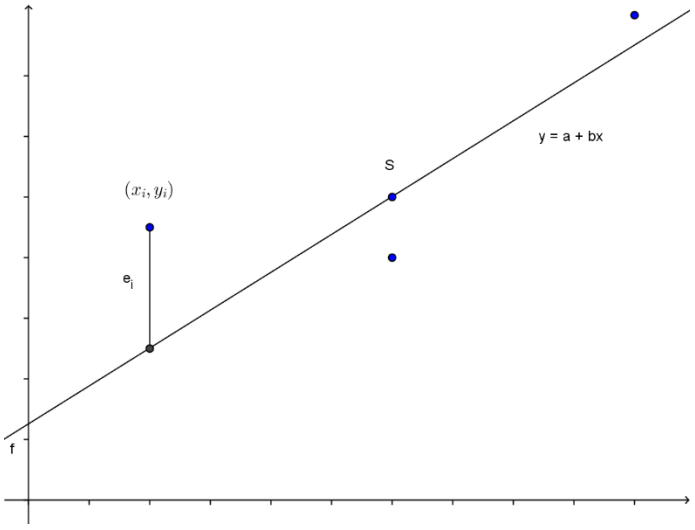
3.3 Die lineare Regression

3.3.1 Berechnung der Ausgleichsgeraden

Bei der Korrelation werden die beiden Grössen x und y gleich behandelt. In vielen Anwendungen ist allerdings die eine Grösse die Zielgrösse y, die andere die Ausgangsgrösse, erklärende Variable x.

Beispiele:

Erklärende Variable	Zielgrösse
Höhe des Wasserfalls	Schwingungsfrequenz
Spraydosen-Verbrauch	Ozongehalt
Geschwindigkeit	Bremsweg
Jahresnote	Maturanote



In vielen Fällen ist „im wesentlichen“ y eine lineare Funktion von x. Damit stellt sich die Frage, wie eine Gerade rechnerisch möglichst gut an eine Stichprobe von Wertepaaren angepasst werden kann.
Die Steigung b (englisch: slope) und der y-Achsenabschnitt (englisch: intercept) sind so zu bestimmen, dass die Abweichungen der Punkte von der Geraden möglichst klein sind. Als Kriterium wird nach Gauss und Legendre (um 1800) die Methode der Kleinsten Quadrate (least squares) verwendet (Maximum likelihood):
Die Steigung b und der y-Achsenabschnitt a der Geraden sind so zu bestimmen, dass die folgende Quadratsumme minimal wird:

$$Q(a,b) = \sum_{i=1}^{n} e_i^2$$

Es kann gezeigt werden, dass die Regressionsgerade durch den „Schwerpunkt“ S $(\bar{x}, \bar{y})$ der Punkte geht.
Die Lösung des Problems vereinfacht sich deshalb, wenn man wie bereits im Abschnitt Korrelation erläutert von den ursprünglichen Daten zu standardisierten Daten (u, v) übergeht. d.h. die Abweichungen der x bzw. y-Werte vom entsprechenden Mittelwert werden in Einheiten der Standardabweichung gemessen.

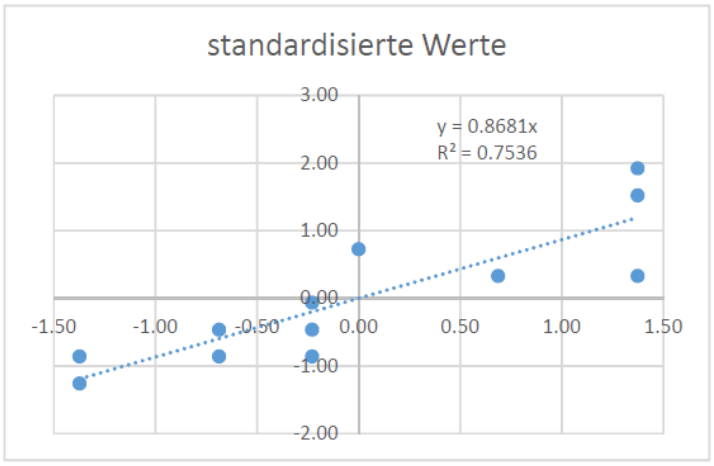
$$u_i = \frac{x_i - \bar{x}}{s_x}$$

und

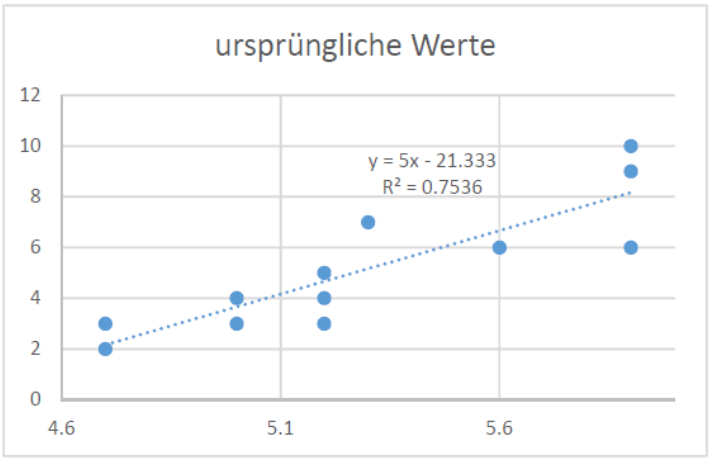
$$v_i = \frac{y_i - \bar{y}}{s_y}$$

Das folgende Beispiel dient der Illustration der Idee

		standardisiert	
x_i	y_i	u_i	v_i
4.7	3	-1.37	-0.86
5	3	-0.69	-0.86
5.2	4	-0.23	-0.46
5.2	5	-0.23	-0.07
5.9	10	1.37	1.92
4.7	2	-1.37	-1.26
5.9	9	1.37	1.52
5.2	3	-0.23	-0.86
5.3	7	0.00	0.73
5.9	6	1.37	0.33
5.6	6	0.69	0.33
5	4	-0.69	-0.46
$\bar{x}$	$\bar{y}$	$\bar{u}_i$	$\bar{v}_i$
5.30	5.17	0.00	0.00
s_x	s_y	s_u	s_v
0.44	2.52	1.00	1.00



Korrelation Wertepaare	standardisiert	
0.868	0.868	
Steigung der Geraden		
5.00	0.868	
Achsenabschnitt		
-21.33	0	
Bestimmtheitsmass	r^2	0.7536
	$\boldsymbol{r^2}$	0.7536



Bemerkung:
In der Abbildung ist ausser dem Quadrat des Korrelationskoeffizienten r_{xy}^2 das erst später definierte Bestimmtheitsmass R^2 angegeben. Im Beispiel stimmt das Quadrat des Korrelationskoeffizienten r_{xy}^2 mit dem Bestimmtheitsmass R^2 überein. In 16) wird gezeigt, dass diese Aussage allgemein gilt.

Wie das Beispiel zeigt, bringt die Standardisierung die folgenden Vereinfachungen:

Die Mittelwerte sind 0:

$$\sum_{i=1}^{n} u_i = n \cdot \bar{u} = 0$$

und

$$\sum_{i=1}^{n} v_i = n \cdot \bar{v} = 0$$

Die Standardabweichungen sind 1 und wegen 2) gilt:

$$(n-1)s_u^2 = n-1 = \sum_{i=1}^{n} u_i^2 - n \cdot \bar{u}^2$$

oder also

$$\boxed{\sum_{i=1}^{n} u_i^2 = n-1} \qquad 8)$$

und entsprechend

$$(n-1)s_v^2 = n-1 = \sum_{i=1}^{n} v_i^2 - n \cdot \bar{v}^2$$

oder also

$$\boxed{\sum_{i=1}^{n} v_i^2 = n-1} \qquad 8')$$

und nach Definition der Korrelation zwischen u_i und v_i nach 7)

$$\boxed{r_{uv} = \frac{1}{n-1}\sum_{i=1}^{n} u_i \cdot v_i} \qquad 9)$$

Da im (u, v)-System die Regressionsgerade durch den Ursprung, den Schwerpunkt geht, kann sie in der Form $v = mu$ angesetzt werden. Die Steigung m ist so zu bestimmen, dass die folgende Quadratsumme minimal wird:

$$Q(m) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (v_i - m \cdot u_i)^2$$

Nach der Kettenregel gilt dann für die 1. Ableitung von Q:

$$Q'(m) = -2\sum_{i=1}^{n} (v_i - m \cdot u_i) \cdot u_i = 0$$

und nach Ausmultiplizieren und Vereinfachen

$$\sum_{i=1}^{n} u_i\, v_i = \mathrm{m} \sum_{i=1}^{n} u_i^2$$

und wegen der Standardisierung nach 9) und 8)

$$(n-1)r_{uv} = (n-1) \cdot m$$

und nach Division durch $(n-1)$

$$m = r_{uv}$$

Für die standardisierten Daten stimmt also die Steigung m der Regressionsgeraden mit der Korrelation r_{uv} überein.

Die Steigung der Regressionsgeraden für die Originaldaten ergibt sich, indem man die Standardisierung rückgängig macht.

$$r_{uv} = m = \frac{v_2 - v_1}{u_2 - u_1} = \frac{\frac{y_2 - \bar{y}}{s_y} - \frac{y_1 - \bar{y}}{s_y}}{\frac{x_2 - \bar{x}}{s_x} - \frac{x_1 - \bar{x}}{s_x}} = \frac{y_2 - y_1}{x_2 - x_1} \cdot \frac{s_x}{s_y} = b \cdot \frac{s_x}{s_y}$$

Löst man diese Gleichung nach b auf, so erhält man für die Steigung der Regressionsgeraden zu den Originaldaten

$$b = r_{uv} \cdot \frac{s_y}{s_x}$$ Steigung der Regressionsgeraden

Da sich bei einer linearen Transformation die Korrelation nicht ändert, gilt $r_{uv} = r_{xy}$ und damit

$$b = r_{xy} \cdot \frac{s_y}{s_x}$$ Steigung der Regressionsgeraden 10)

Da für die Korrelation gemäss 5´) gilt:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

erhält man

$$b = r_{xy} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$

oder

$$b = \frac{s_{xy}}{s_x^2}$$ oder $$s_{xy} = b \cdot s_x^2$$ 11)

Damit ist auch der y-Achsenabschnitt bestimmt, denn da der „Schwerpunkt“ S $(\bar{x}, \bar{y})$ auf der Regressionsgeraden liegt gilt:

$\bar{y} = a + b\bar{x}$ oder
$a = \bar{y} - b\bar{x}$

Zusammenfassung:
Die Regressionsgerade hat die Steigung

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$ 12)

und den y-Achsenabschnitt

$$a = \bar{y} - b\bar{x}$$ 13)

Bemerkung:
Ist die Kettenregel nicht bekannt, so kann die Steigung m auch folgendermassen bestimmt werden:
Multipliziert man bei Q(m) die Summanden nach der binomischen Formel aus, so erhält man eine in m quadratische Funktion der Form
$Q(m) = a \cdot m^2 + bm + c$
Der Graph von Q ist eine nach oben geöffnete Parabel, deren Scheitelabszisse gleich dem ersten Summanden in der quadratischen Auflösungsformel ist. Für das Minimum von Q ergibt sich damit
$m = -\frac{b}{2a}$
in Übereinstimmung mit 11)

3.3.2 Das Bestimmtheitsmass

Die Abweichungen $e_i = y_i - \hat{y}_i$ zwischen den beobachteten Werten y_i und den durch die Regressionsgerade vorhergesagten Werten $\hat{y}_i$ heissen Residuen. Mit ihnen kann punktweise überprüft werden, wie gut das Modell mit den Daten übereinstimmt. Es fehlt aber ein Mass um die Güte des Regressionsmodells insgesamt beurteilen zu können. Ein solches Mass kann über die sogenannte Streuungszerlegung erhalten werden.

Dazu führt man die folgenden Summen ein:
SQT ("**S**um of s**q**uares **T**otal") ist die Gesamtstreuung,

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SQE („**S**um of s**q**uares **E**xplained") ist durch die Regression erklärte Streuung

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SQR (**S**um of s**q**uares **R**esidual) ist durch die Rest- oder Residualstreuung erklärte Streuung.

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Für diese drei Summen gilt die folgende
Streuungszerlegung
$SQT = SQE + SQR$ oder $SQR = SQT - SQE$
also

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad 14)$$

wobei für die y-Werte der Regressionsgeraden gilt: $\hat{y}_i = a + b \cdot x_i$

Beweis:
Aus
$y_i - \hat{y}_i = y_i - (a + bx_i)$
und
$a = \bar{y} - b\bar{x}$
folgt
$y_i - \hat{y}_i = y_i - (\bar{y} - b\bar{x} + bx_i) = (y_i - \bar{y}) - b(x_i - \bar{x})$

und damit nach dem binomischen Lehrsatz

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) + b^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Da wegen 11) gilt:

$$\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = b\sum_{i=1}^{n}(x_i - \bar{x})^2$$

erhält man schliesslich

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

SQR = SQT - SQE
oder auch

$$b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SQE = SQT - SQR
Wegen
$(\hat{y}_i - \bar{y})^2 = (bx_i + a - \bar{y})^2 = \left(bx_i + a - (b\bar{x} + a)\right)^2 = b^2(x_i - \bar{x})^2$
gilt also wie behauptet
$SQR = SQT - SQE$

Als Masszahl für die Güte des Modells wird das Bestimmtheitsmass R^2 (auch Determinationskoeffizient) verwendet. Das Bestimmtheitsmass gibt an, welchen Anteil der Gesamtstreuung die durch die Regression erklärte Streuung ausmacht.

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad 15)$$

Aus der Definition folgt: $0 \le R^2 \le 1$

Spezialfälle:
Für R = 0 ist das Modell denkbar schlecht
Für R = 1 liegen die Datenpunkte auf einer Geraden.

Daraus folgt nun, dass der quadrierte Korrelationskoeffizient r_{xy}^2 gerade mit dem Bestimmtheitsmass übereinstimmt, denn nach 5´) und 11) gilt:

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = \frac{s_{xy}^2 \cdot s_x^2}{s_x^4 \cdot s_y^2} = \frac{s_{xy}^2}{s_x^4} \cdot \frac{s_x^2}{s_y^2} = b^2 \cdot \frac{s_x^2}{s_y^2} = \frac{b^2 \cdot s_x^2}{s_y^2} = \frac{SQE}{SQT} = R^2$$

Damit gilt allgemein:

$$r_{xy}^2 = R^2 \qquad 16)$$

Beispiel:

Wasserfälle

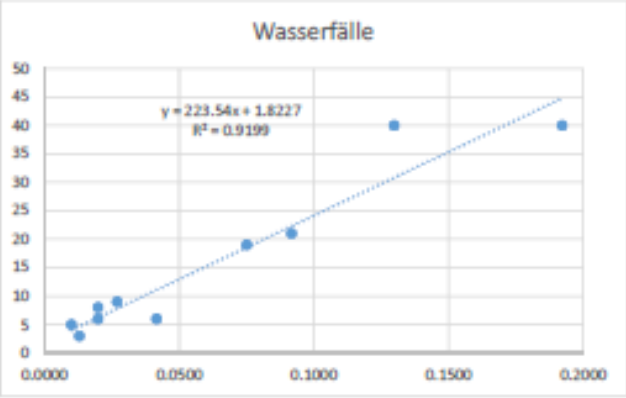
i	Hoehe [m] h_i	Kehrwert $x_i = {}^1/_{z_i}$	Frequenz [Hz] y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$	$e_i = \hat{y}_i - \bar{y}$	e_i^2	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	100	0.0100	5	0.00	25	0.050	4.1	-0.9	0.89	-10.7	114.49
2	77	0.0130	3	0.00	9	0.039	4.7	1.7	2.98	-12.7	161.29
3	50	0.0200	6	0.00	36	0.120	6.3	0.3	0.09	-9.7	94.09
4	50	0.0200	8	0.00	64	0.160	6.3	-1.7	2.91	-7.7	59.29
5	37	0.0270	9	0.00	81	0.243	7.9	-1.1	1.29	-6.7	44.89
6	24	0.0417	6	0.00	36	0.250	11.1	5.1	26.39	-9.7	94.09
7	13.3	0.0752	19	0.01	361	1.429	18.6	-0.4	0.14	3.3	10.89
8	10.9	0.0917	21	0.01	441	1.927	22.3	1.3	1.77	5.3	28.09
9	7.7	0.1299	40	0.02	1600	5.195	30.9	-9.1	83.64	24.3	590.49
10	5.2	0.1923	40	0.04	1600	7.692	44.8	4.8	23.15	24.3	590.49
	375.1	0.6208	157	0.07	4253	17.104		0.0	143.2	0.0	1788.1
	37.51	0.0621					$\bar{x}$		SQR		SQT
			15.7				$\bar{y}$				
				0.00							SQE
				0.06			s_x				1644.9
					198.68						
					14.10		s_y				

Steigung b der Regressionsgeraden: 0.82 s_{xy}

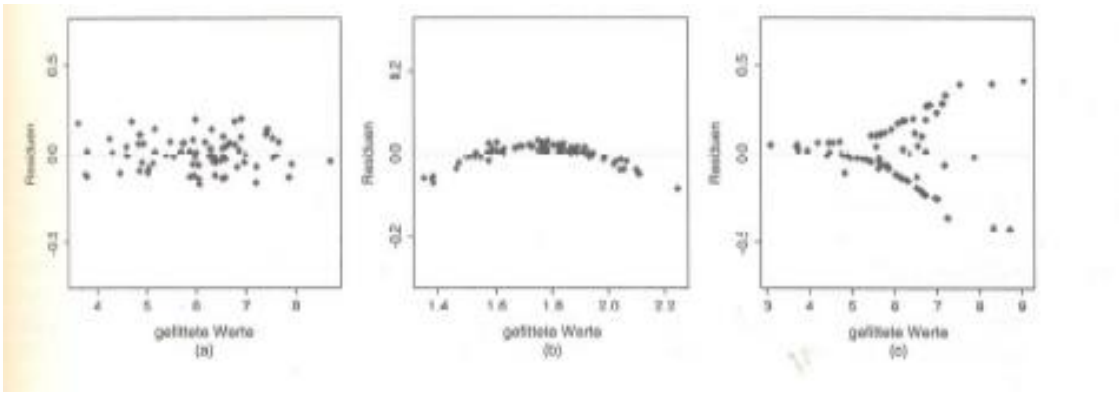
223.54 223.54

Achsenabschnitt: 0.959 r_{xy}

1.8227: 0.9199 r_{xy}^2 R^2 0.9199



Bei der Beurteilung der Güte des Modells sollte aber nicht allein das Bestimmtheitsmass, sondern auch das Muster der Residuen betrachtet werden.
Das Verhalten in der Abbildung a) ist ideal, denn die Residuen schwanken unsystematisch und sie sind nahezu 0.
Im Gegensatz dazu verändern sich die Residuen bei b) und c) je nach der Einflussgrösse. Eine nicht lineare Abhängigkeit wird durch das Modell nicht beschrieben.



Beispiele aus verschiedenen Anwendungsgebieten

a) Finanzwirtschaft: Das CAP-Modell (Quelle: Fahrmeier)

Bei Anwendungen im Finanzbereich wird die Steigung b der Regressionsgeraden mit β und der y-Achsenabschnitt a mit α bezeichnet.
Das „Capital Asset Pricing Model“ wird verwendet, um das Risiko von verschiedenen Aktien zu vergleichen. Dazu wird auf der Basis von Finanzmarktdaten über einen längeren Zeitraum für jede Aktie der sogenannte Beta-Faktor bestimmt. Mit diesem Faktor kann das Risiko dieser Aktie gemessen am Risiko des Gesamtmarkts beurteilt werden. Beispielsweise bedeutet ein Beta-Faktor grösser als eins, dass das Risiko der Aktie grösser als das Marktrisiko ist.
Es interessiert der Zusammenhang zwischen der Einflussgrösse x (Marktrisiko minus Zins) und dem abhängigen Merkmal y (Rendite der Aktie minus Zins).

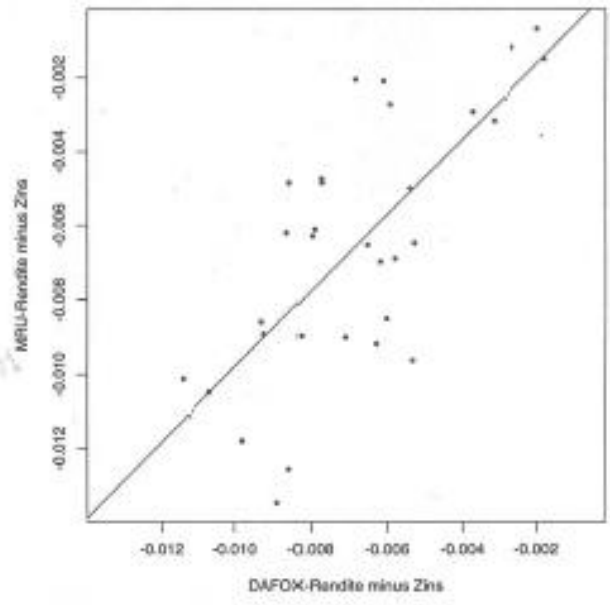


ABBILDUNG 3.20: Streudiagramm und Ausgleichsgerade für das CAP-Modell

Im abgebildeten Beispiel ergibt sich aufgrund der Daten die Regressionsgerade
$y = 0.0004 + 1.0216x$
Das Ergebnis bedeutet, dass im Mittel
bei einer Änderung der Marktrendite x um ein Prozent sich die Rendite y der Aktie um 1.02% also wegen $\beta \approx 1.02 \approx 1$ etwa auch um ein Prozent verändert.
Die Residuen enthalten die Renditekomponenten der Aktie, die sich nicht durch den Markt erklären lassen. Diese Aktie hat also ein mit dem Markt vergleichbares Risiko. Das Bestimmtheitsmass $R^2 = 0.5$ bedeutet, dass 50% der Renditeschwankungen der Aktie durch Schwankungen des Marktes bedingt sind und 1 - 0.5 d.h. 50% titelspezifischer Natur sind.
Der Korrelationskoeffizient r_{xy} zwischen den Marktrendite und Rendite der Aktie beträgt
$\sqrt{0.5} \approx 71\%$

Das Residuenmuster zeigt keine besonderen Auffälligkeiten

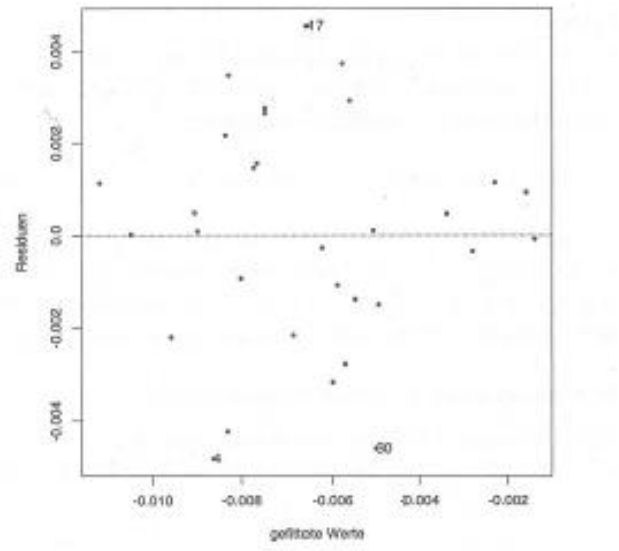


ABBILDUNG 3.21: Residualplot für das CAP-Modell

b) Physik: Die Zustandsgleichung für ideale Gase

Bei konstantem Druck p ist das Volumen V eines Gases nach dem Gesetz von Gay-Lussac proportional zur absoluten Temperatur T d.h. es gilt:

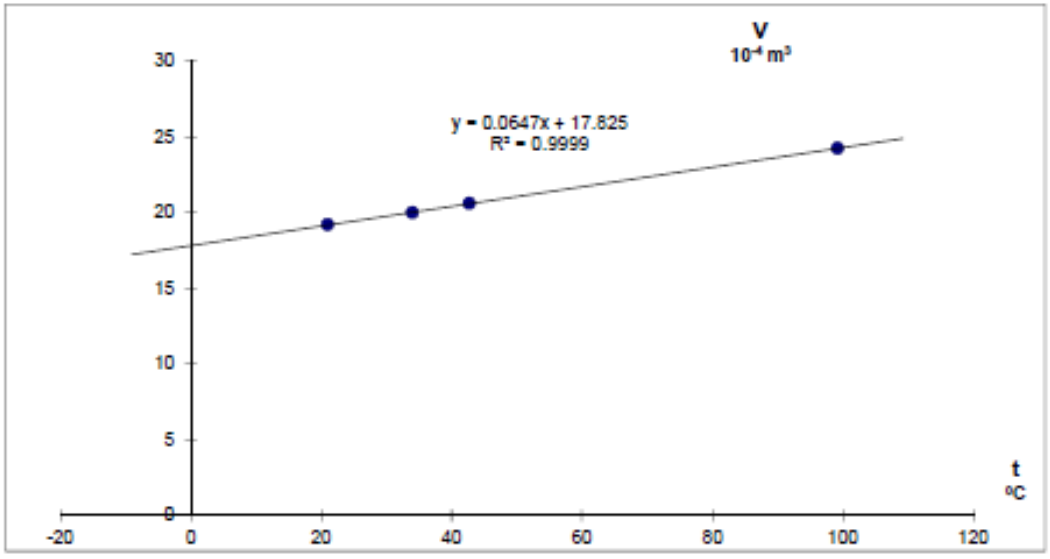
$V = cT$
Die Proportionalitätskonstante c ist vom konstanten Druck und von der vorliegenden Gasmenge abhängig. Misst man die Temperatur in °C, dann ist das Volumen also eine lineare Funktion der Temperatur

$\vartheta = T - T_n$
wobei T_n die sogenannte Normtemperatur 273.15 bedeutet.
Für das Volumen gilt damit
$V(\vartheta) = V_0 + b \cdot \vartheta$

Dabei ist V_0 das Volumen bei 0°C (der y-Achsenabschnitt der Regressionsgeraden) und b die Steigung der Regressionsgeraden. Die folgenden Daten stammen von einem Versuch im Physikunterricht an der KSZ (ac)

Volumenausdehnung **2C 26.6.1994**

Nr.	Temp. T ^0C	V cm^3	$V(t) = V_0+\gamma V_0 t$			
			$m = \gamma V_0$	V_0	γ	T_0
1	21.0	19.20				
2	34.0	20.00	0.06474905	17.8252282	0.00363	-275.3
3	42.7	20.60				
4	99.2	24.25				



Die bekannte Rechnung ergibt für b = 0.06474905 und für a = V_0 = 17.8252282.
Die Ausgleichsgerade schneidet die Temperaturachse bei -275.3°C, dem absoluten Nullpunkt der Temperatur. Ein genauerer Wert ist -273.15°C.

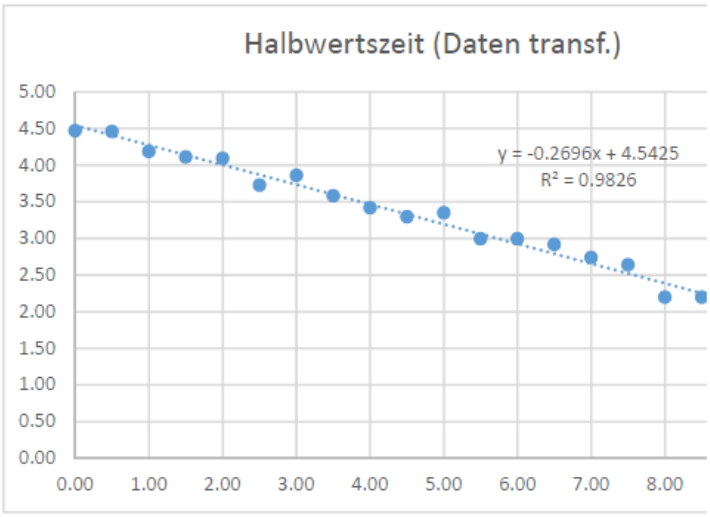
Besteht zwischen Einflussgrösse und Zielgrösse kein linearer Zusammenhang, dann kann dies eine geeignete Transformation der Daten ermöglichen.

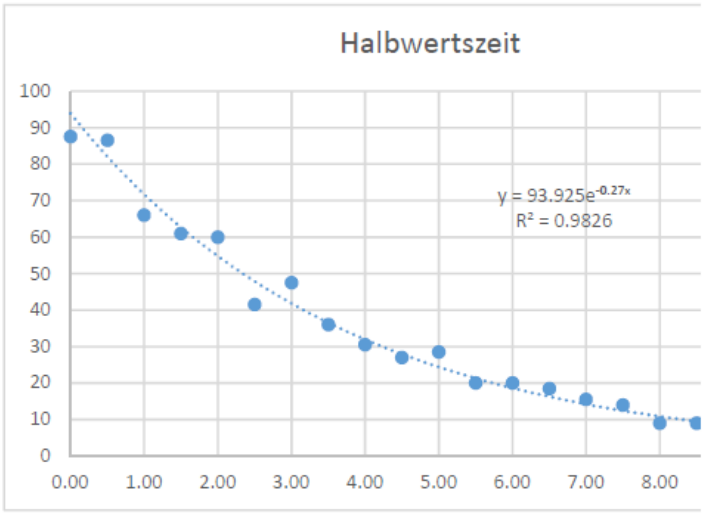
Wachstumsgesetz	Einflussgrösse	Zielgrösse
Exponentialfunktion	ln(Einflussgrösse)	Zielgrösse
Potenzfunktion	ln(Einflussgrösse)	ln(Zielgrösse)

c) Halbwertszeit von Ba-137 (SPAM-Maturaaufgabe an der AKSA 2007)
Zur Bestimmung der Halbwertszeit von Ba-137 wurde im Abstand von 30 Sekunden die Impulsrate pro Sekunde gemessen. Dabei ergaben sich die in der Tabelle ermittelten Werte:

Halbwertszeit

Zeit [min] t	Impulsrate r $[s^{-1}]$	$ln(r)$
0.00	87.5	4.47
0.50	86.5	4.46
1.00	66	4.19
1.50	61	4.11
2.00	60	4.09
2.50	41.5	3.73
3.00	47.5	3.86
3.50	36	3.58
4.00	30.5	3.42
4.50	27	3.30
5.00	28.5	3.35
5.50	20	3.00
6.00	20	3.00
6.50	18.5	2.92
7.00	15.5	2.74
7.50	14	2.64
8.00	9	2.20
8.50	9	2.20
9.00	7.5	2.01





Für den radioaktiven Zerfall gilt das Gesetz:

$N(t) = N(0) \cdot e^{-kt}$
Logarithmiert man diese Gleichung, dann gilt:

$ln\big(N(t)\big) = ln\big(N(0)\big) - kt$

d.h. die logarithmierten Funktionswerte sind eine lineare Funktion der Zeit. Die Ausgleichsgerade für die Punkte $\big(t, ln\big(N(t)\big)\big)$ hat die Steigung
b = -0.2695772 und den y-Achsenabschnitt a = 4.5424927
Wegen $ln\big(N(0)\big) = a$ und und $-k = b$ folgt
$N(0) = e^{4.5425} \approx 93.925$
und damit das Zerfallsgesetz
$N(t) = N(0) \cdot e^{-kt} \approx 93.925 \cdot e^{-0.26958 \cdot t}$

Die Halbwertszeit τ, die auch grafisch bestimmt werden kann, ergibt sich wegen
$N(\tau) = \frac{1}{2} \cdot N(0) = N(0) \cdot e^{-k\tau}$
nach Division durch N(0)
$\frac{1}{2} = e^{-k\tau}$
zu
$\tau = \frac{ln2}{k} \approx 2.57\ min$

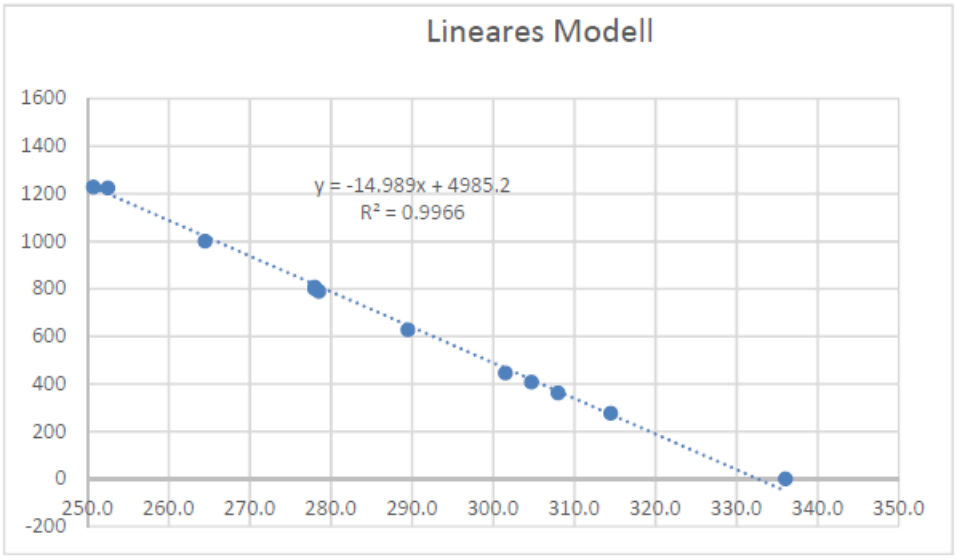
d) Barometerdruck in verschiedenen Höhenlagen
Quelle: Lambert, Beiträge zum Gebrauch der Mathematik, Theil 1, Berlin 1765
Es wurden die Höhen über Meer verschiedener Berge im Languedoc, der Auvergne und der Provence gemessen.
Längenmass: 1 Toise entspricht 1.494 m
Barometer: Linie
Lambert interessierte sich für die Form der Abhängigkeit des Drucks von der Höhe. Im folgenden Beispiel interessiert umgekehrt die für einen Höhenmesser wichtige Abhängigkeit der Höhe vom Druck. In den ersten drei Spalten der Tabelle sind die von Lambert erfassten Werte aufgeführt.

Wählt man für die Daten ein lineares Modell, dann ist die Höhe des Berges eine lineare Funktion des Barometerdrucks p oder
$h = b \cdot p + a$

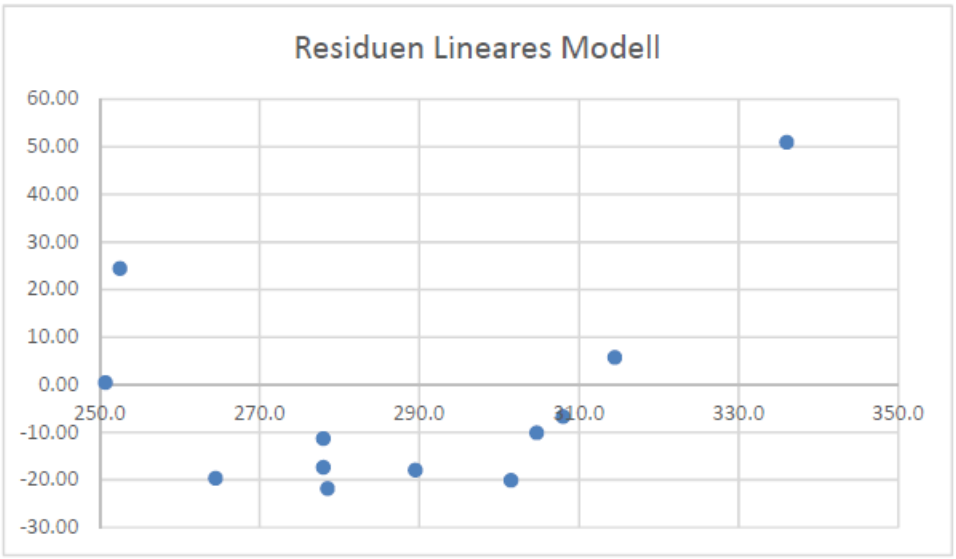
Barometerdruck in verschiedenen Höhenlagen

Lineares Modell			**a**	**b**
			4985.2	**-14.989**
Ort	Druck p	Höhe h	erwartet	Residuen
Meereshöhe	336.0	0	-50.99	50.99
Clairet	314.5	277	271.27	5.73
Rodez	308	362	368.69	-6.69
Massane	304.7	408	418.16	-10.16
Rupeyroux	301.5	446	466.12	-20.12
Bugarac	289.5	628	645.98	-17.98
Puy du Dôme	278.5	789	810.86	-21.86
La Coste	278	807	818.35	-11.35
La Courlande	278	801	818.35	-17.35
Mont d'Or	264.5	1001	1020.70	-19.70
St. Barthélémy	252.5	1225	1200.56	24.44
Mousset	250.7	1228	1227.54	0.46
Canigou	240.5	1424	1380.42	43.58

Die zugehörigen Punkte sind in der folgenden Abbildung dargestellt:



Wie in den bisherigen Beispielen können der y-Achsenabschnitt a und die Steigung b bestimmt werden mit dem folgenden Ergebnis: a = -14.989 und b = 4985.2.
Mit diesen Werten können die nach dem Modell geschätzten Höhen bestimmt werden, die in der vierten Spalte der Tabelle dargestellt sind. Der Skizze entnimmt man, dass das Modell die Daten scheinbar gut beschreibt. Zudem ist auch das Bestimmtheitsmass sehr hoch.
Berücksichtigt man hingegen die in der fünften Spalte aufgeführten Residuen, so zeigen diese ein auffälliges Muster:



Masszahlen oder grafische Darstellungen sollten nicht überbewertet werden. Es ist auch zu überprüfen, ob die Residuen - wie in diesem Beispiel- allenfalls von der Einflussgrösse abhängen.

Nun wird die die Abhängigkeit des Drucks von der Höhe ein exponentielles Modell gewählt:
$p(h) = p_0 \cdot e^{-kh}$

Logarithmiert man diese Gleichung so erhält man:
$ln\big(p(h)\big) = ln(p_0) - kh$

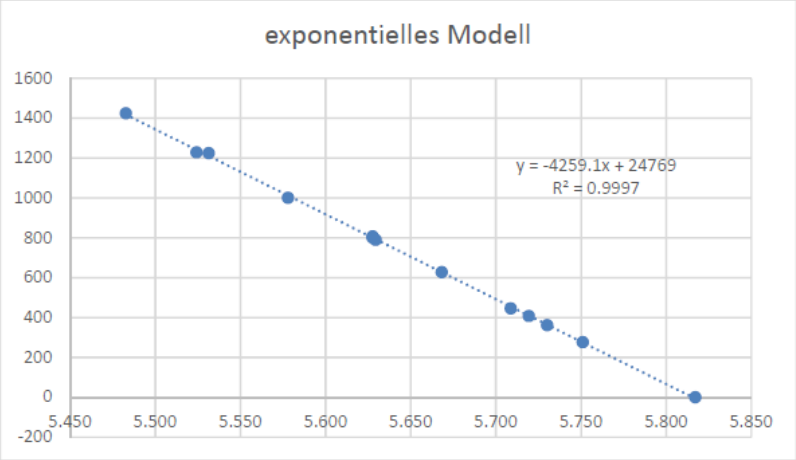
Oder nach h aufgelöst:
$ln\big(p(h)\big) - ln(p_0) = -kh$
bzw.
$ln(p_0) - ln\big(p(h)\big) = kh$
oder schliesslich nach Division durch k
$h = \frac{1}{k} ln(p_0) - \frac{1}{k} ln\big(p(h)\big)$
Bei dieser Transformation der Daten ist also die Höhe eine lineare Funktion von $ln\big(p(h)\big)$

Barometerdruck in verschiedenen Höhenlagen

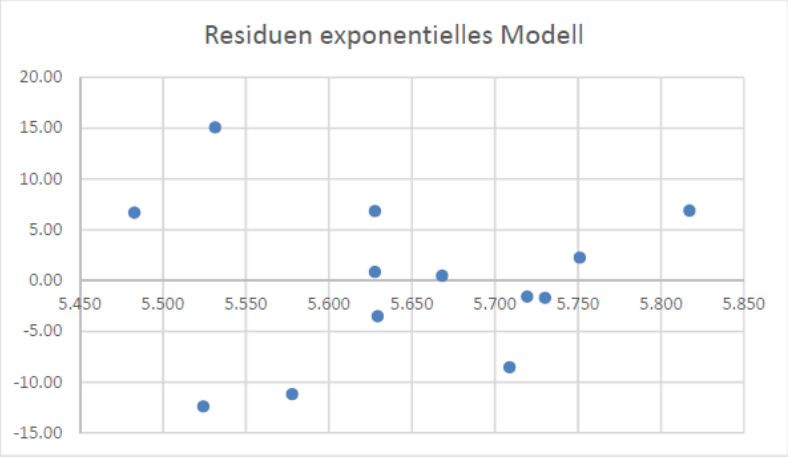
Exponentielles Modell

Ort	Druck p	ln (p)	Höhe h	erwartet	Residuen
Meereshöhe	336.0	5.817	0	-6.88	6.88
Clairet	314.5	5.751	277	274.76	2.24
Rodez	308	5.730	362	363.71	-1.71
Massane	304.7	5.719	408	409.59	-1.59
Rupeyroux	301.5	5.709	446	454.55	-8.55
Bugarac	289.5	5.668	628	627.54	0.46
Puy du Dôme	278.5	5.629	789	792.52	-3.52
La Coste	278	5.628	807	800.17	6.83
La Courlande	278	5.628	801	800.17	0.83
Mont d'Or	264.5	5.578	1001	1012.19	-11.19
St. Barthélémy	252.5	5.531	1225	1209.94	15.06
Mousset	250.7	5.524	1228	1240.41	-12.41
Canigou	240.5	5.483	1424	1417.32	6.68

In der Abbildung sind die Punkte $\left(ln\left(p(h)\right), h\right)$ dargestellt:



Für die Ausgleichsgerade ergeben sich die folgenden Werte: a = 24768.7 und b = -4259.1



Das Muster der Residuen zeigt keine Auffälligkeiten.

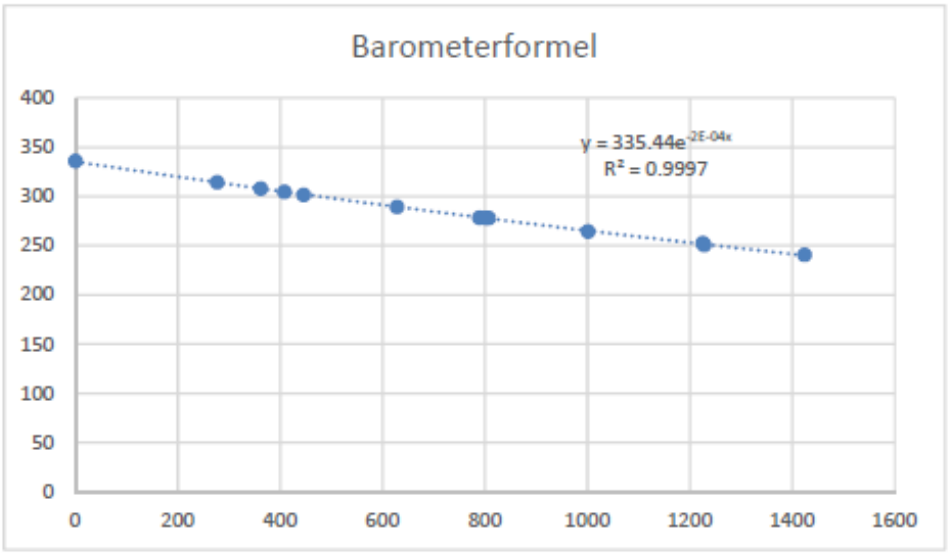
Die Parameter der Barometerformel ergeben sich aus der Steigung und dem y-Achsenabschnitt der Ausgleichsgeraden.

$-\frac{1}{k} = b$ oder $k = -\frac{1}{b} \approx 0.00023479$

$\frac{1}{k} ln(p_0) = a$ oder $ln(p_0) = k \cdot a \approx 5.8155$ und $p_0 \approx 335.458$

Damit gilt mit den erwähnten Einheiten die folgende Barometerformel:

$p(h) = p_0 \cdot e^{-kh} = 335.46 \cdot e^{-0.00023479 \cdot h}$



e) Weltrekorde in der Leichtathletik

In der folgenden Tabelle sind die Weltrekorde der Leichtathletik dargestellt (Stand: August 2014).
Trägt man in einem Koordinatensystem sowohl die Distanz (in Meter) und die Zeit (in Sekunden) logarithmisch ab, so liegen die Punkte annähernd auf einer Geraden. Dies bedeutet, dass die Beziehung zwischen Distanz und Weltrekordzeit ungefähr durch eine Potenzfunktion gegeben ist.

Weltrekorde

Distanz [m]	Zeit [s]	ln(Distanz) x_i	ln(Zeit) y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$
100.00	9.58	4.6052	2.260	21.21	5.10614	10.406	2.2
200.00	19.19	5.2983	2.954	28.07	8.72842	15.653	3.0
400.00	43.18	5.9915	3.765	35.90	14.1781	22.560	3.8
800.00	100.92	6.6846	4.614	44.68	21.292	30.845	4.6
1000.00	131.96	6.9078	4.882	47.72	23.8388	33.727	4.8
1500.00	206.00	7.3132	5.328	53.48	28.3863	38.964	5.3
3000.00	440.67	8.0064	6.088	64.10	37.0674	48.745	6.1
5000.00	757.35	8.5172	6.630	72.54	43.9546	56.468	6.6
10000.00	1577.54	9.2103	7.364	84.83	54.2229	67.821	7.4
		62.5344	43.8859	452.5366	236.7746	325.1898	
		6.9483					$\bar{x}$
			4.876210135				$\bar{y}$
				2.25			
				1.50			s_x
					2.85		
					1.69		s_y
Steigung b der Regressionsgeraden						2.53	s_{xy}
1.12358878398963	1.12						
Achsenabschnitt a						0.9997	r_{xy}
-2.93078943200053						0.9993	r_{xy}^2

Logarithmiert man die Gleichung einer Potenzfunktion
$t = c \cdot x^b$
so erhält man
$ln(t) = ln(c) + b \cdot ln(x)$
Die Punkte P$(ln(x), ln(t))$ liegen also auf einer Geraden mit der Steigung b und dem y-Achsenabschnitt $a = ln(c)$. Damit ist $c = e^a$.
Die Punkte (x, t) liegen also annähernd auf dem Graphen der Potenzfunktion

$t = 0.05334 \cdot x^{1.1236}$

Obwohl das Bestimmtheitsmass sehr gross ist, beschreibt das Modell die Daten nicht besonders gut. Die Daten weichen in zwei Fällen von den Weltrekorden um mehr als 5% ab und die Residuen zeigen ein auffälliges Muster.

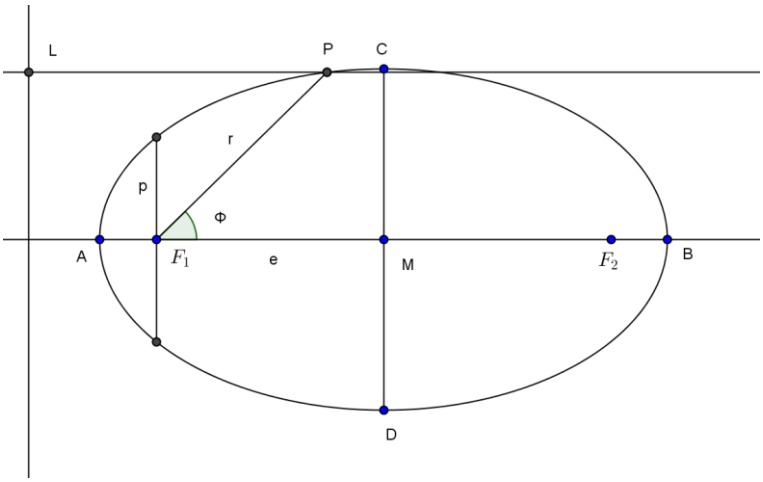
f) Der Komet Tentax

Der 1968 entdeckte Komet Tentax bewegt sich im Sonnensystem. In einem geeigneten Polarkoordinatensystem (r, φ) wurden folgende Positionen des Kometen beobachtet:

φ°	φ (rad)	r
48.0	0.8	2.70
67.0	1.2	2.00
83.0	1.4	1.61
108.0	1.9	1.20
126.0	2.2	1.02

Falls die Einflüsse der Planeten vernachlässigt werden, bewegt sich der Komet nach dem Keplerschen Gesetz auf einer elliptischen oder hyperbolischen Bahn mit der Gleichung

$$r = \frac{p}{1 - e \cdot cos(\varphi)}$$



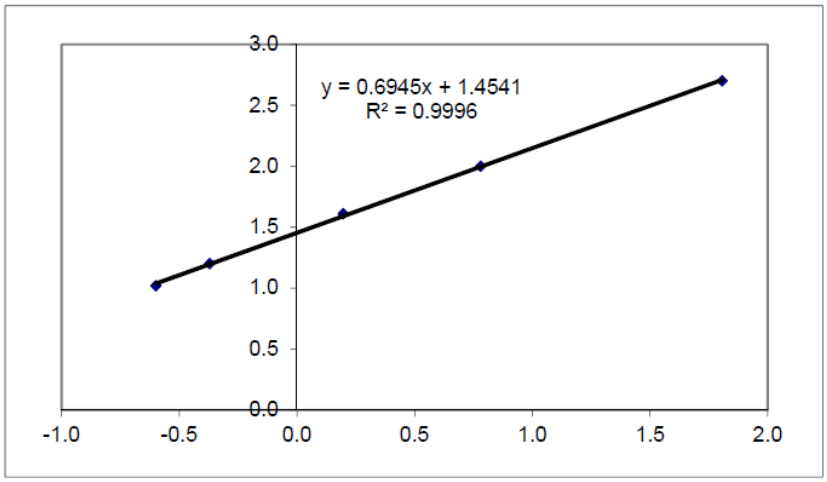
Der Parameter p und die lineare Exzentrizität e können aus den Daten geschätzt werden. Dazu wird die Kegelschnittgleichung als lineare Gleichung umgeschrieben.
Wird
$r \cdot \left(1 - e \cdot cos(\varphi)\right) = r - e \cdot r \cdot cos(\varphi) = p$
nach r aufgelöst
$r = p + r \cdot cos(\varphi) \cdot e$

so folgt mit $x = r \cdot cos(\varphi)$ und $y = r$ die lineare Gleichung
$y = p + e \cdot x$

Die lineare Exzentrizität beträgt e ≈ 0.6945 und der Parameter p ≈ 1.4541

Wegen e < 1 bewegt sich also Tentax auf einer elliptischen Bahn mit dem Parameter p

φ°	φ (rad)	r	x=$r \cdot cos(\varphi)$	y = r
48.0	0.8	2.70	1.807	2.70
67.0	1.2	2.00	0.781	2.00
83.0	1.4	1.61	0.196	1.61
108.0	1.9	1.20	-0.371	1.20
126.0	2.2	1.02	-0.600	1.02



Übungsaufgabe:
J. Kepler (1571 – 1630) hat zu Beginn des 17. Jahrhunderts die Bewegungen der Planetenbahnen untersucht und seine berühmten „Gesetze“ entdeckt. Die Planetenbahnen sind Ellipsen (mit der Sonne im Brennpunkt) und der Verbindungsstrahl Sonne – Erde überstreicht in gleichen Zeiten gleiche Flächen. Das dritte Keplersche Gesetz stellt einen Zusammenhang her zwischen der Halbachse a der Ellipse und der Umlaufszeit T her. T ist eine Potenzfunktion von a

$T = \alpha \cdot a^{\beta}$

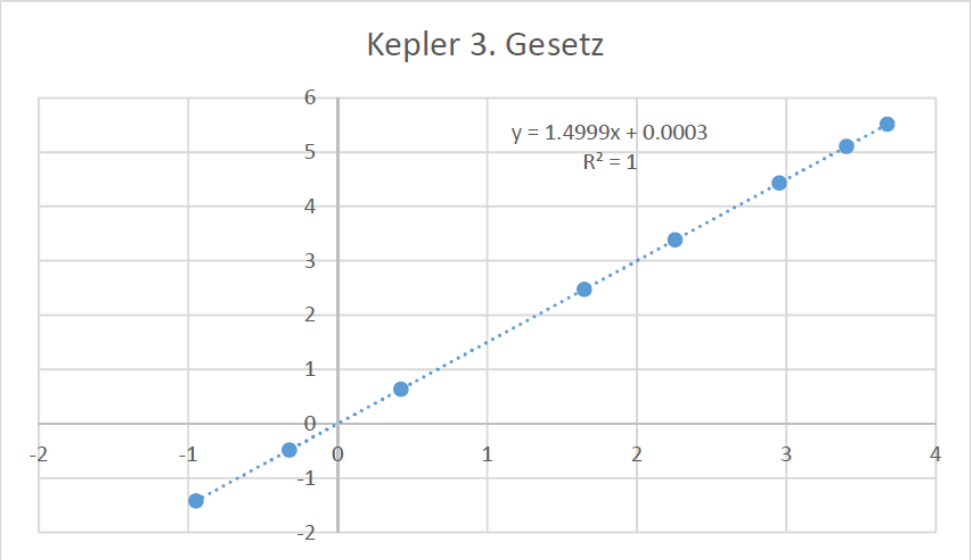
a in astronomischen Einheiten, T in Erdjahren.
Es wurden folgende Daten gemessen (Datenquelle nicht bekannt):

Planet	a	T
Merkur	0.387	0.241
Venus	0.723	0.615
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.458
Uranus	19.182	84.014
Neptun	30.058	164.793
Pluto	39.44	247.7

Wegen

$ln(T) = ln(\alpha) + \beta \cdot ln(a)$

besteht zwischen $ln(T)$ und $ln(a)$ wie die folgende Abbildung zeigt eine lineare Beziehung:



Die Ausgleichsgerade hat die Steigung $\beta \approx {}^{3}/_{2}$ und den y-Achsenabschnitt $0.0003 \approx \ln(\alpha)$. Es gilt also näherungsweise wegen $\alpha \approx 1$

$$T \approx a^{\frac{3}{2}}$$

oder

$$\frac{T^2}{a^3} \approx konstant$$

Heute ist bekannt, dass die Keplerschen Gesetze die Bewegungen der Planeten nur näherungsweise beschreiben.