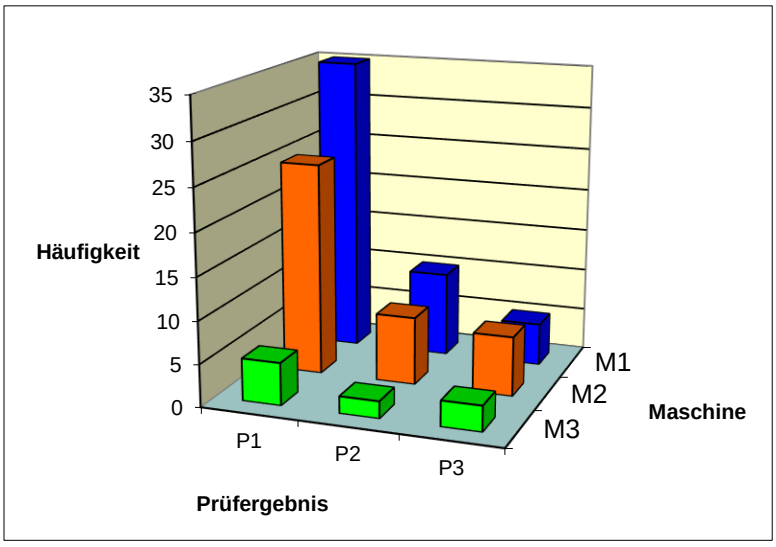
Def.: Die gleichzeitige Beschreibung von zwei bzw. mehreren Variablen heißt **bivariate** bzw. **multivariate Verteilung**. Eine zweidimensionale Häufigkeitstabelle heißt **Kontingenztabelle** oder **Kontingenztafel** oder **Kreuztabelle**.

Beispiel 1: Drei Maschinen 1, 2 und 3 stellen Teile her, die einwandfrei sind (1), kleine Mängel haben (2) oder unbrauchbar sind (3). Eine Überprüfung ergab folgende Kontingenztafel:

	Prüfergebnis			
Maschine	***1***	***2***	***3***	$\sum$
1	35	10	5	50
2	25	8	7	40
3	5	2	3	10
$\sum$	65	20	15	**n = 100**



Und nun zu den beiden **bedingten Verteilungen**:

1. Bedingte Verteilung für die Zeilen (Zeilenvergleich)

	Prüfergebnis			
Maschine	***1***	***2***	***3***	Zeilensum-men
1	$35/50 = 0,7$	$10/50 = 0,2$	$5/50 = 0,1$	1
2	$25/40 = 0,625$	$8/40 = 0,2$	$7/40 = 0,175$	1
3	$5/10 = 0,5$	$2/10 = 0,2$	$3/10 = 0,3$	1
	$65/100 = 0,65$	$20/100 = 0,2$	$15/100 = 0,15$	1

Interpretation:
Zu Feld (1,1): 70% der von Maschine 1 hergestellten Teile sind einwandfrei.
Zu Feld (1,2): 20% der von Maschine 1 hergestellten Teile sind mit kleinen Mängeln.
Zu Feld (1,3): 10% der von Maschine 1 hergestellten Teile sind unbrauchbar.
Zu Feld (2,1): 62,5% der von Maschine 2 hergestellten Teile sind einwandfrei.
…
Zu Feld (4,1): 65% aller hergestellten Teile sind einwandfrei.
Zu Feld (4,2): 20% aller hergestellten Teile sind mit kleinen Mängeln.
Zu Feld (4,3): 15% aller hergestellten Teile sind unbrauchbar.

Welche Maschine würden Sie sich zulegen? Natürlich Maschine 1, da sie mit $0,7 = 70\%$ die meisten einwandfreien Teile herstellt.

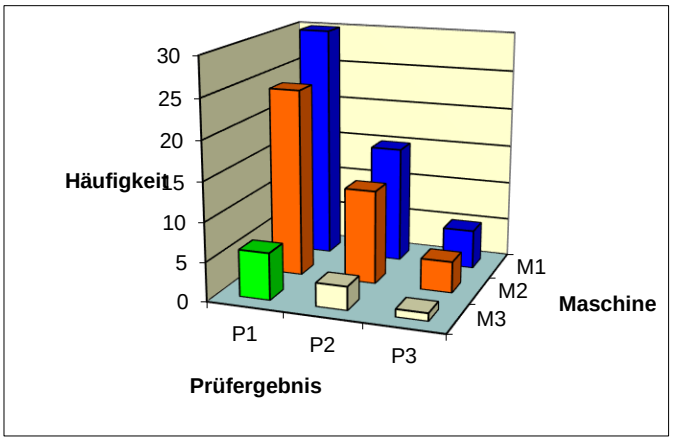
2. Bedingte Verteilung für die Spalten (Spaltenvergleich)

	Prüfergebnis			
Maschine	***1***	***2***	***3***	
1	$35/65 \approx 0,538$	$10/20 = 0,5$	$5/15 \approx 0,333$	$50/100 = 0,5$
2	$25/65 \approx 0,385$	$8/20 = 0,4$	$7/15 \approx 0,467$	$40/100 = 0,4$
3	$5/65 \approx 0,077$	$2/20 = 0,1$	$3/15 = 0,2$	$10/100 = 0,1$
Spaltensum-men	1	1	1	1

Interpretation:
Zu Feld (1,1): ≈ 53,8% der einwandfreien Teile stammen vom Maschine 1. (Wieder optimal)
Zu Feld (2,1): ≈ 38,5% der einwandfreien Teile stammen von Maschine 2.
Zu Feld (3,1): ≈ 7,7% der einwandfreien Teile stammen von Maschine 3.
Zu Feld (1,2): 50% der Teile mit kleinen Mängeln stammen von Maschine 1.
... ...
Zu Feld (1,4): 50% aller hergestellten Teile stammen von Maschine 1.
Zu Feld (2,4): 40% aller hergestellten Teile stammen von Maschine 2.
Zu Feld (3,4): 10% aller hergestellten Teile stammen von Maschine 3.

Beispiel 2: Drei Maschinen 1, 2 und 3 stellen Teile her, die einwandfrei sind (1), kleine Mängel haben (2) oder unbrauchbar sind (3). Eine Überprüfung ergab folgende Kontingenztafel:

	Prüfergebnis			
Maschine	***1***	***2***	***3***	$\sum$
1	30	15	5	50
2	24	12	4	40
3	6	3	1	10
$\sum$	60	30	10	**n = 100**



Und nun zu den beiden **bedingten Verteilungen**:

1. Bedingte Verteilung für die Zeilen (Zeilenvergleich)

	Prüfergebnis			
Maschine	***1***	***2***	***3***	Zeilensum-men
1	$30/50 = 0,6$	$15/50 = 0,3$	$5/50 = 0,1$	1
2	$24/40 = 0,6$	$12/40 = 0,3$	$4/40 = 0,1$	1
3	$6/10 = 0,6$	$3/10 = 0,3$	$1/10 = 0,1$	1
	$60/100 = 0,6$	$30/100 = 0,3$	$10/100 = 0,1$	1

Wir sehen, dass die jeweiligen Prozentwerte **unabhängig** vom Maschinentyp sind.

2. Bedingte Verteilung für die Spalten (Spaltenvergleich)

	Prüfergebnis			
Maschine	***1***	***2***	***3***	
1	$30/60 = 0,5$	$15/30 = 0,5$	$5/10 = 0,5$	$50/100 = 0,5$
2	$24/60 = 0,4$	$12/30 = 0,4$	$4/10 = 0,4$	$40/100 = 0,4$
3	$6/60 = 0,1$	$3/30 = 0,1$	$1/10 = 0,1$	$10/100 = 0,1$
Spaltensummen	1	1	1	1

Wir sehen wieder, dass die jeweiligen Prozentwerte **unabhängig** vom Prüfergebnis sind.

Wie muss eine Tabelle aufgebaut sein, damit ihre **beiden Variablen unabhängig** sind?

	Prüfergebnis			
Maschine	***1***	***2***	***3***	$\sum$
1	$n_{\bullet 1} \cdot n_{1\bullet} / n$	$n_{\bullet 2} \cdot n_{1\bullet} / n$	$n_{\bullet 3} \cdot n_{1\bullet} / n$	$n_{1\bullet}$
2	$n_{\bullet 1} \cdot n_{2\bullet} / n$	$n_{\bullet 2} \cdot n_{2\bullet} / n$	$n_{\bullet 3} \cdot n_{2\bullet} / n$	$n_{2\bullet}$
3	$n_{\bullet 1} \cdot n_{3\bullet} / n$	$n_{\bullet 2} \cdot n_{3\bullet} / n$	$n_{\bullet 3} \cdot n_{3\bullet} / n$	$n_{3\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	**n**

Maßzahlen des rechnerischen Zusammenhangs

Es interessiert, wie sehr die Werte einer Kontingenztafel von den Idealwerten bei Unabhängigkeit der beiden Variablen abweichen.
Es bezeichne h_{ij} die absolute Häufigkeit des Wertes in Zeile Nr. i und Spalte Nr. j.

Weiter sei h_{ij}^e die ‚**e**rwartete' absolute Häufigkeit des Wertes in Zeile Nr. i und Spalte Nr. j für den Fall der Unabhängigkeit der beiden Variablen. Das ‚e' kann auch für ‚**e**xpected' stehen.

Dann nennt man $\boxed{\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(h_{ij} - h_{ij}^e\right)^2}{h_{ij}^e}}$ den $\chi^2 - $Koeffizienten .

Beispiel: Wir verwenden die Daten von Beispiel 1 oben:

	Prüfergebnis			
Maschine	***1***	***2***	***3***	$\sum$
1	35	10	5	50
2	25	8	7	40
3	5	2	3	10
$\sum$	65	20	15	**n = 100**

In jedes der 9 Felder notieren wir die 4 Werte:

h_{ij}	h_{ij}^e
$(h_{ij} - h_{ij}^e)^2$	$\frac{(h_{ij} - h_{ij}^e)^2}{h_{ij}^e}$

Zur Wiederholung: $h_{11}^e = \frac{65 \cdot 50}{100} = 32,5$, $h_{12}^e = \frac{20 \cdot 50}{100} = 10$, usw.

	Prüfergebnis						
Maschine	***1***		***2***		***3***		$\sum$
1	35	32,5	10	10	5	7,5	50
	6,25	5/26	0	0	6,25	5/6	
2	25	26	8	8	7	6	40
	1	1/26	0	0	1	1/6	
3	5	6,5	2	2	3	1,5	10
	2,25	9/26	0	0	2,25	3/2	
$\sum$	65		20		15		**n = 100**

Und damit $\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\left(h_{ij}-h_{ij}^e\right)^2}{h_{ij}^e} = \left(\frac{5}{26}+0+\frac{5}{6}\right)+\left(\frac{1}{26}+0+\frac{1}{6}\right)+\left(\frac{9}{26}+0+\frac{3}{2}\right) = \frac{40}{13} \approx 3{,}08$.

Der Nachteil des χ^2-Koeffizienten besteht darin, dass z.B. bei Verdoppelung aller Werte auch χ^2 verdoppelt wird. Man führt deshalb den **Kontingenzkoeffizienten** nach **Karl Pearson** (brit. Mathematiker, 1857-1936) $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$ ein, für den sicher $0 \le C < 1$ gilt. Dabei ist n die Anzahl der Messpunkte.

Von *diesem* C kann man zeigen, dass der größte Wert gleich $C_{max} = \sqrt{\frac{M-1}{M}}$ beträgt, wobei M der kleinere Wert von Zeilen- und Spaltenanzahl ist. Dann gilt also $0 \le C \le C_{max}$.

Für den **korrigierten C-Koeffizient** $C^* = \frac{C}{C_{max}}$ gilt somit $\boxed{C^* = \sqrt{\frac{M}{M-1}\cdot\frac{\chi^2}{\chi^2+n}}}$ mit $\boxed{0 \le C^* \le 1}$.

In unserem Beispiel ist $C^* = \sqrt{\frac{3}{2}\cdot\frac{40/13}{40/13+100}} = \sqrt{\frac{3}{67}} \approx 0{,}212$.

Was bedeuten $C^* = 0$ und $C^* = 1$?

Beispiel 1: Wenn alle absoluten Häufigkeiten $h_{ij} = h_{ij}^e$ gleich den Werten bei **Unabhängigkeit** sind, dann wird $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(h_{ij}-h_{ij}^e\right)^2}{h_{ij}^e} = 0$ und damit auch $\mathbf{C^* = 0}$.

Beispiel 2: Gegeben sei die folgende Kontingenztafel. Die Variablen X und Y sind sicher abhängig voneinander.

	x_1	x_2	$\sum$
y_1	1	0	1
y_2	0	1	1
$\sum$	1	1	n = 2

	x_1		x_2		$\sum$
y_1	1	0,5	0	0,5	1
	0,25	0,5	0,25	0,5	
y_2	0	0,5	1	0,5	1
	0,25	0,5	0,25	0,5	
$\sum$	1		1		n = 2

Es folgt $\chi^2 = 0{,}5+0{,}5+0{,}5+0{,}5 = 2$.

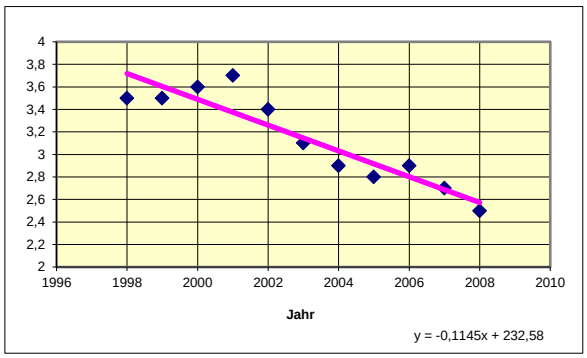
Außerdem ist $C = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{2}{2+2}} = \sqrt{\frac{1}{2}} \approx 0{,}707$ und $C^* = \frac{C}{C_{max}} = \sqrt{\frac{2}{2-1}\cdot\frac{2}{2+2}} = 1$, was die vollständige Abhängigkeit belegt.

Für den Fall der **Abhängigkeit** zweier Variablen gilt $\mathbf{C^* = 1}$.

Die lineare Regression

Beispiel: Der Prozentsatz der nicht versetzten Gymnasiasten in Baden-Württemberg ist nicht jedes Jahr gleich.

Jahr	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Pro-zent	3,5	3,5	3,6	3,7	3,4	3,1	2,9	2,8	2,9	2,7	2,5



Excel liefert eine Ausgleichsgerade der Gleichung $y = -0{,}1145x + 232.58$. Man erkennt, dass der Prozentsatz der nicht versetzten Gymnasiasten mit der Zeit abnimmt.
Wie erhält man diese Gerade?

Gegeben sind n Wertepaare (x_i / y_i), $i = 1, ..., n$. Gesucht ist derjenige lineare Zusammenhang $\boxed{y = a + b \cdot x}$, der allen n Wertepaaren ‚am nächsten kommt'.

Wenn man y_i durch $a + b \cdot x_i$ ersetzt, so begeht man den Fehler $e_i = y_i - (a + b \cdot x_i) = y_i - a - b \cdot x_i$. Diese Fehler können positiv oder negativ sein. Die Bedingung für die beste Ausgleichsgerade kann also nicht sein, dass die Summe aller Fehler e_i Null sein soll. Denn dann ist es möglich, dass sich große positive Fehler durch große negative Fehler aufheben können.

Deshalb muss die Bedingung lauten: $\sum_{i=1}^{n} |e_i|$ minimal oder $\sum_{i=1}^{n} e_i^2$ minimal.

Es ist einfacher mit Quadraten als mit Beträgen zu rechnen. Deshalb verwendet man die zweite Bedingung.
Man nennt sie auch **Methode der kleinsten Quadrate**.

Bestimme die beiden Zahlen a und b so, dass die Summe $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2$ minimal wird.

Die notwendige Bedingung für die Bestimmung eines Minimums ist, dass die ersten partiellen Ableitungen nach a und nach b Null sind.

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2 = -2 \cdot \sum_{i=1}^{n} (y_i - a - b \cdot x_i) = 0 \quad \text{und} \quad \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2 = -2 \cdot \sum_{i=1}^{n} (y_i - a - b \cdot x_i) \cdot x_i = 0 .$$

Umgeformt $a \cdot n + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ und $a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i$. Und nach Division durch n folgt

$$a + b \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{und} \quad a \cdot \frac{1}{n} \sum_{i=1}^{n} x_i + b \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i \quad \text{bzw.} \quad a + b \cdot \bar{x} = \bar{y} \quad \text{und} \quad a \cdot \bar{x} + b \cdot \overline{x^2} = \overline{x \cdot y} .$$

Daraus folgt

$$b = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{XY}}{s_X^2} \quad \text{und} \quad a = \bar{y} - b \cdot \bar{x}$$

mit der **Kovarianz** $s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y} = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$

und der **Varianz** $s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$

Zurück zu obigem Beispiel:

i	Jahr x_i	Prozent y_i	x_i^2	y_i^2	$x_i \cdot y_i$		$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	1998	3,5	3992004	12,25	6993		-5	25	-0,2
2	1999	3,5	3996001	12,25	6996,5		-4	16	-0,16
3	2000	3,6	4000000	12,96	7200		-3	9	-0,42
4	2001	3,7	4004001	13,69	7403,7		-2	4	-0,48
5	2002	3,4	4008004	11,56	6806,8		-1	1	0,06
6	2003	3,1	4012009	9,61	6209,3		0	0	0
7	2004	2,9	4016016	8,41	5811,6		1	1	-0,56
8	2005	2,8	4020025	7,84	5614		2	4	-1,32
9	2006	2,9	4024036	8,41	5817,4		3	9	-1,68
10	2007	2,7	4028049	7,29	5418,9		4	16	-3,04
11	2008	2,5	4032064	6,25	5020		5	25	-4,8
$\sum$	**22033**	**34,6**	**44132209**	**110,52**	**69291,2**		**0**	**110**	**-12,6**

Damit wird $\bar{x} = \frac{22033}{11} = 2003$, $\bar{y} = \frac{34,6}{11} \approx 3,145$,

also $s_{XY} = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = \frac{1}{11} \cdot 69291,2 - 2003 \cdot \frac{34,6}{11} = -\frac{63}{55} \approx -1,145$

und $s_X^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{11} \cdot 44132209 - 2003^2 = 10$.

Ergebnis: $b = \frac{s_{xy}}{s_x^2} = \frac{-63/55}{10} = -\frac{63}{550} \approx -0,1145$ und $a = \bar{y} - b \cdot \bar{x} = \frac{34,6}{11} + \frac{63}{550} \cdot 2003 = 232,58$, so dass die Regressionsgerade die Gleichung $y = 232,58 - 0,1145 \cdot x$ besitzt.

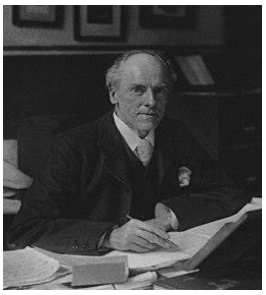
Nach dem zweiten Formelsystem folgt $s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{11} \cdot (-12,6) = -\frac{63}{55}$ und

$s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{11} \cdot 110 = 10$ wie oben.

An der Formel für die Kovarianz $s_{XY} = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$ erkennt man, dass ihr Wert von der Stärke des Zusammenhangs zwischen x und y und von der Größe der x_i und y_i abhängt. Als Maß für die Güte der Regressionsgeraden führt man den **Bravais-Pearson-Korrelationskoeffizienten** $\boxed{r = \frac{s_{XY}}{s_X \cdot s_Y}}$ ein.

Auguste Bravais, französischer Physiker (1811-1863)

Karl Pearson, britischer Statistiker (1857-1936)



Satz: a. Der Quotient $r = \frac{s_{XY}}{s_X \cdot s_Y}$ liegt zwischen -1 und $+1$.

b. Bei $r = 1$ liegen alle Punkte auf einer aufsteigenden Geraden.
Bei $r = -1$ liegen alle Punkte auf einer absteigenden Geraden.
Bei $r \approx 0$ besteht kein linearer Zusammenhang zwischen den beiden Merkmalen.

Bew.: a. $r = \dfrac{s_{XY}}{s_X \cdot s_Y} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ lässt sich als Skalarprodukt der beiden Vektoren $\begin{pmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{pmatrix}$ und $\begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$ interpretieren. r ist damit gleich einem Cosinus und liegt folglich zwischen -1 und $+1$.

b. Die n Punkte (x_i / y_i) sollen nun alle auf der Geraden $y = a + b\cdot x$ liegen, so dass $y_i = a + b\cdot x_i$ für alle i gilt. Wegen $a = \bar{y} - b\cdot\bar{x}$ gilt auch $\bar{y} = a + b\cdot\bar{x}$. Dies ergibt

$$r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(a+b\cdot x_i-a-b\cdot\bar{x}\right)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(a+b\cdot x_i-a-b\cdot\bar{x})^2}} =$$

$$\frac{b\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{b^2\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{b}{|b|} = \begin{cases} 1 & \text{für } b>0 \\ -1 & \text{für } b<0 \end{cases}$$. Und b ist die Steigung der Regressionsgeraden.

Im Fall $r = \dfrac{s_{XY}}{s_X \cdot s_Y} = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{s_X \cdot s_Y} = 0$ gilt $\overline{xy} = \bar{x}\cdot\bar{y}$, d.h. die Variablen x und y sind unabhängig voneinander.

In obigem Beispiel ist $s_{XY} = -\dfrac{63}{55}$, $s_X = \sqrt{10}$ und $s_Y^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \dfrac{1}{11}\cdot 110{,}52 - \left(\dfrac{34{,}6}{11}\right)^2 = \dfrac{464}{3025}$, also $s_Y = \sqrt{\dfrac{464}{3025}}$. Und damit $r = \dfrac{-63/55}{\sqrt{10}\cdot\sqrt{464/3025}} \approx -0{,}925$, d.h. die Punkte bestimmen recht gut eine abfallende Gerade, wie man auch am Schaubild erkennen kann.

Noch ein Beispiel: Man bestimme die Regressionsgerade durch die drei gegebenen Punkte.

x_i	-2	0	1
y_i	-3	1	3

Daraus folgt

$$\overline{x\cdot y} = \frac{1}{3}\left((-2)\cdot(-3)+0\cdot 1+1\cdot 3\right) = \frac{1}{3}(6+0+3) = 3$$

$$\bar{x} = \frac{1}{3}\left((-2)+0+1\right) = -\frac{1}{3}, \qquad \bar{y} = \frac{1}{3}\left((-3)+1+3\right) = \frac{1}{3}$$

$$\overline{x^2} = \frac{1}{3}\left((-2)^2+0^2+1^2\right) = \frac{5}{3}, \qquad \overline{y^2} = \frac{1}{3}\left((-3)^2+1^2+3^2\right) = \frac{19}{3}.$$

Also $b = \dfrac{\overline{x\cdot y} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{3-\left(-\frac{1}{3}\right)\cdot\frac{1}{3}}{\frac{5}{3}-\frac{1}{9}} = \dfrac{\frac{28}{9}}{\frac{14}{9}} = \dfrac{28}{9}\cdot\dfrac{9}{14} = 2$ und $a = \bar{y} - b\cdot\bar{x} = \dfrac{1}{3} - 2\cdot\left(-\dfrac{1}{3}\right) = 1$.

Ergebnis: Die Regressionsgerade hat die Gleichung $y = 1 + 2x$.

Der Korrelationskoeffizient wird $r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{3 - \left(-\frac{1}{3}\right) \cdot \frac{1}{3}}{\sqrt{\frac{5}{3} - \frac{1}{9}} \cdot \sqrt{\frac{5}{3} - \frac{1}{9}}} = \frac{\frac{28}{9}}{\sqrt{\frac{14}{9} \cdot \frac{56}{9}}} = \frac{\frac{28}{9}}{\frac{28}{9}} = 1$.

Die drei Punkte liegen sogar exakt auf dieser aufsteigenden Geraden.

Die exponentielle Regression

Beispiel: Die Höhe y in cm des Bierschaums in einem Glas wird über vier Minuten ($0 \le x \le 4$) gemessen.

x in min	0	1	2	3	4
y in cm	5,0	2,2	1,0	0,5	0,2

Es ist bekannt, dass ein exponentieller Zusammenhang $y = c \cdot e^{d \cdot x}$ besteht. Wie bestimmt man c und d?
Durch Logarithmieren folgt $\ln y = \ln c + d \cdot x = a + b \cdot x$ also ein linearer Zusammenhang zwischen $z = \ln y$ und x. Somit können wir obige Formeln verwenden.

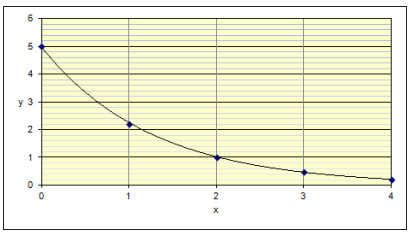
i	x_i	y_i	$z_i = \ln y_i$	x_i^2	z_i^2	$x_i \cdot z_i$		$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (z_i - \bar{z})$
0	0	5	1,609	0	2,590	0,000		-2	4	-3,180
1	1	2,2	0,788	1	0,622	0,788		-1	1	-0,769
2	2	1,0	0,000	4	0,000	0,000		0	0	0
3	3	0,5	-0,693	9	0,480	-2,079		1	1	-0,712
4	4	0,2	-1,609	16	2,590	-6,436		2	4	-3,256
$\sum$	10	---	0,095	30	6,282	-7,727		0	10	-7,917

Damit wird $\bar{x} = \frac{10}{5} = 2$, $\bar{z} = \frac{0,095}{5} = 0,019$,

also $s_{XZ} = \overline{x \cdot z} - \bar{x} \cdot \bar{z} = \frac{1}{5} \cdot (-7,727) - 2 \cdot 0,019 = -1,5834$

und $s_X^2 = \overline{x^2} - \bar{x}^2 = \frac{1}{5} \cdot 30 - 2^2 = 2$.

Ergebnis: $b = \frac{s_{xz}}{s_x^2} = \frac{-1,5834}{2} = -0,7917$ und $a = \bar{z} - b \cdot \bar{x} = 0,019 + 0,7917 \cdot 2 = 1,6024$, so dass die Regressionsgerade die Gleichung $z = 1,6024 - 0,7917 \cdot x$ besitzt. Und mit $z = \ln y$ folgt $\ln y = 1,6024 - 0,7917 \cdot x$. d.h.
$y = e^{1,6024 - 0,7917 \cdot x} = e^{1,6024} \cdot e^{-0,7917 \cdot x} = 4,96 \cdot e^{-0,79 \cdot x}$, also
$y = 4,96 \cdot e^{-0,79 \cdot x}$, siehe Schaubild.



Nach dem zweiten Formelsystem folgt
$s_{XZ} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (z_i - \bar{z}) = \frac{1}{5} \cdot (-7,917) = -1,5834$ und
$s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{5} \cdot 10 = 2$ wie oben.

Mit $s_Z^2 = \frac{1}{n} \sum_{i=1}^{n} z_i^2 - \bar{z}^2 = \frac{1}{5} \cdot 6,282 - \left(\frac{0,095}{5}\right)^2 = 1,256$ folgt der Korrelationskoeffizient r zu

$r = \frac{s_{XZ}}{s_X \cdot s_Z} = \frac{-1,5834}{\sqrt{2} \cdot \sqrt{1,256}} = -0,999$, so dass die Punkte (x_i / z_i) so gut wie perfekt auf einer absteigenden Gerade liegen.

Die quadratische Regression

Gegeben sind n Wertepaare (x_i / y_i), $i = 1, ..., n$. Gesucht ist derjenige Zusammenhang $\boxed{y = a \cdot x^2 + b \cdot x + c}$, der allen n Wertepaaren ‚am nächsten kommt'.

Nach der **Methode der kleinsten Quadrate** sind die drei Zahlen a, b und c so zu bestimmen, dass die Summe der Fehlerquadrate $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right)^2$ minimal wird.

Die notwendige Bedingung für die Bestimmung eines Minimums ist, dass die ersten Ableitungen nach a, b und c Null sind.

$\frac{\partial}{\partial a} \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right)^2 = -2 \cdot \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) \cdot x_i^2 = 0$ und

$\frac{\partial}{\partial b} \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right)^2 = -2 \cdot \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) \cdot x_i = 0$ und

$\frac{\partial}{\partial c} \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right)^2 = -2 \cdot \sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) = 0$.

Nach Division von $\sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) \cdot x_i^2 = 0$ durch n folgt $\overline{y \cdot x^2} - a \cdot \overline{x^4} - b \cdot \overline{x^3} - c \cdot \overline{x^2} = 0$.

Nach Division von $\sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) \cdot x_i = 0$ durch n folgt $\overline{y \cdot x} - a \cdot \overline{x^3} - b \cdot \overline{x^2} - c \cdot \overline{x} = 0$.

Nach Division von $\sum_{i=1}^{n} \left(y_i - a \cdot x_i^2 - b \cdot x_i - c\right) = 0$ durch n folgt $\overline{y} - a \cdot \overline{x^2} - b \cdot \overline{x} - c = 0$.

Dies ist ein lineares Gleichungssystem für die drei Unbekannten a, b und c.

Die multiple lineare Regression

Die Größe y soll von k Variablen $x_1, x_2, ..., x_k$ mit $k \in \mathbb{N}$ linear abhängen, so dass $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k$ gilt.

Beispiel: y sei das Ergebnis einer Klausur, x_1 die Vorbereitungszeit der Studenten, x_2 die Länge der Klausur, x_3 die Körpertemperatur des Studenten, usw.

Wir beschränken uns auf den Fall $k = 2$. Jede Messung hat dann die Form (x_1, x_2, y).

Nr.	1	2	3	⋯	n
x_1	x_{11}	x_{12}	x_{13}	⋯	x_{1n}
x_2	x_{21}	x_{22}	x_{23}	⋯	x_{2n}
y	y_1	y_2	y_3	⋯	y_n

Dabei muss n mindesten 3 sein, da drei Unbekannte β_0, β_1 und β_2 bestimmt werden sollen.

Nach der Methode der kleinsten Quadrate müssen diese β_0, β_1 und β_2 so bestimmt werden, dass die Summe $\left(y_1 - \beta_0 - \beta_1 \cdot x_{11} - \beta_2 \cdot x_{21}\right)^2 + \left(y_2 - \beta_0 - \beta_1 \cdot x_{12} - \beta_2 \cdot x_{22}\right)^2 + ... + \left(y_1 - \beta_0 - \beta_1 \cdot x_{1n} - \beta_2 \cdot x_{2n}\right)^2$, also

$\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right)^2$ minimal wird. Durch partielles Ableiten folgt

$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right)^2 = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right) = 0$

$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right)^2 = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right) \cdot x_{1i} = 0$

$\frac{\partial}{\partial \beta_2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right)^2 = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \cdot x_{1i} - \beta_2 \cdot x_{2i}\right) \cdot x_{2i} = 0$.

Nach Division durch n folgen drei linearen Gleichungen für β_0, β_1 und β_2.

$\bar{y} - \beta_0 - \beta_1 \cdot \bar{x_1} - \beta_2 \cdot \bar{x_2} = 0$

$\overline{y \cdot x_1} - \beta_0 \cdot \bar{x_1} - \beta_1 \cdot \overline{x_1^2} - \beta_2 \cdot \overline{x_1 \cdot x_2} = 0$

$\overline{y \cdot x_2} - \beta_0 \cdot \bar{x_2} - \beta_1 \cdot \overline{x_1 \cdot x_2} - \beta_2 \cdot \overline{x_2^2} = 0$.

Die logistische Regression

Bei der logistischen Regression wird untersucht, mit welcher Wahrscheinlichkeit ein bestimmtes Ereignis y unter den Voraussetzungen x eintritt.

Beispiele:
x sei das Alter einer Person, y sei der Kauf eines bestimmten Buches. Dabei interessiert den Händler, mit welcher Wahrscheinlichkeit eine x Jahre alte Person dieses Buch kauft.
x sei die täglich konsumierte Alkoholmenge einer Person, y sei das Auftreten einer bestimmten Krankheit. Dabei interessiert den Mediziner, mit welcher Wahrscheinlichkeit ein x-Trinker diese Krankheit bekommt.
x sei das Alter einer Person, y sei das Auftreten einer bestimmten Krankheit. Dabei interessiert den Mediziner, mit welcher Wahrscheinlichkeit eine Person von x Jahren diese Krankheit bekommt.
x sei die tägliche Stundenzahl, die ein Student Statistik lernt, y das Bestehen der Statistik-Klausur. Wieder interessiert, mit welcher Wahrscheinlichkeit ein Student, der täglich x Stunden Statistik lernt, die Klausur besteht.

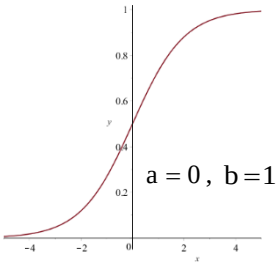
Die allgemeine logistische Funktion lautet $f(x) = \frac{1}{1+e^{-(a+b\cdot x)}}$ für $x \in \mathbb{R}$ und mit den Parametern $a, b \in \mathbb{R}$. Man erkennt, dass $0 < f(x) < 1$ gilt und die Grenzwerte $\lim\limits_{x \to \pm\infty} f(x)$ die Werte 0 oder 1 haben. Wegen

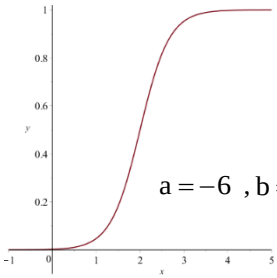
$f'(x) = b \cdot \frac{e^{-(a+b\cdot x)}}{\left(1+e^{-(a+b\cdot x)}\right)^2}$ ist f für $b > 0$ streng monoton steigend und für $b < 0$ streng monoton fallend.

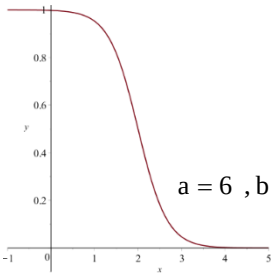
Aus $f''(x) = \frac{b^2 \cdot e^{-(a+b\cdot x)} \cdot \left(e^{-(a+b\cdot x)} - 1\right)}{\left(1+e^{-(a+b\cdot x)}\right)^3}$ ergibt sich der Wendepunkt $W\left(-\frac{a}{b} \middle| \frac{1}{2}\right)$, zu dem das Schaubild von f

symmetrisch ist. Denn $g(x) = f\left(x - \frac{a}{b}\right) - \frac{1}{2} = \frac{1}{2} \cdot \frac{1-e^{-b\cdot x}}{1+e^{-b\cdot x}}$ ist wegen $g(-x) = -g(x)$ symmetrisch zum Ursprung.

Drei Beispiele:







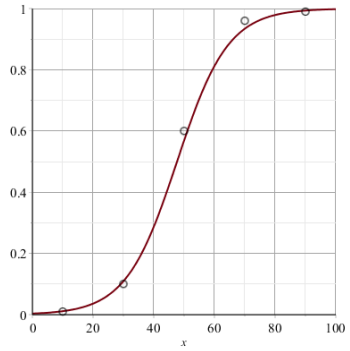
Nun zum Verfahren:

Gegeben sei ein Datensatz (x_i / y_i) für $i = 1, ..., n$. Wie kann man damit die Parameter a und b schätzen?

Aus $y = \frac{1}{1+e^{-(a+b\cdot x)}}$ folgt $1 + e^{-(a+b\cdot x)} = \frac{1}{y}$, folglich $e^{-(a+b\cdot x)} = \frac{1-y}{y}$ und $-(a + b \cdot x) = \ln\left(\frac{1-y}{y}\right)$, sodass schließlich $a + b \cdot x = \ln\left(\frac{y}{1-y}\right)$ gilt. Und darauf kann man die lineare Regression anwenden.

Zahlenbeispiel: Es soll untersucht werden, wie eine bestimmte Krankheit vom Alter der Person abhängt. Dazu wurden die untersuchten Personen in fünf Altersgruppen $0 \le x < 20$, $20 \le x < 40$, $40 \le x < 60$, $60 \le x < 80$ und $80 \le x < 100$ eingeteilt und jeweils notiert, mit welcher relativen Häufigkeit y diese Krankheit bei ihnen auftrat. Es ergab sich das folgende Ergebnis: x bezeichnet den Altersmittelwert der jeweiligen Gruppe.

x	10	30	50	70	90
y	0,01	0,10	0,60	0,96	0,99
$\frac{y}{1-y}$	0,0101	0,1111	1,5000	24,000	99,000
$z = \ln\left(\frac{y}{1-y}\right)$	–4,959	–2,197	0,405	3,178	4,595



Aus diesen Daten folgt $\bar{x} = 50$, $\bar{z} = 0,277$, $\overline{x^2} = 3300$, $\overline{x \cdot z} = 108,886$ und daraus $b = \frac{s_{XZ}}{s_X^2} = \frac{\overline{x \cdot z} - \bar{x} \cdot \bar{z}}{\overline{x^2} - \bar{x}^2} = 0,119$ und $a = \bar{z} - b \cdot \bar{x} = -5,661$. Insgesamt ergibt sich $y = \frac{1}{1 + e^{-(-5,661 + b0,119x)}}$; siehe Schaubild.

Zusatz: Der Quotient $\frac{y}{1-y}$ aus der Wahrscheinlichkeit y und der Gegenwahrscheinlichkeit $1-y$ heißt „odds“ (Chance). Sein Logarithmus $\ln\left(\frac{y}{1-y}\right)$ heißt auch der „logit“.

Zum Beispiel beträgt das Odds in der Gruppe $x = 50$ gerade $1,5$, d.h. dass hier die Chance diese Krankheit zu haben $1,5 = 3:2$ beträgt, dass es also 1,5-mal so wahrscheinlich ist diese Krankheit zu haben als nicht zu haben. In der Gruppe $x = 70$ beträgt die Chance bereits $24 = 24:1$.

Und weiter: Wie groß ist der Quotient aus dem Odds für $x = 70$ durch den Odds für $x = 50$, also die relative Chance? Dieser Quotient heißt das „odds ratio“. Er beträgt $\frac{24}{1,5} = 16$. Das heißt, die Chance (das Risiko) als Siebzigjähriger die Krankheit zu bekommen ist 16-mal höher als bei einem Fünfzigjährigen.
Der umgekehrte Quotient beträgt $\frac{1,5}{24} = \frac{1}{16} = 0,0625$. Das heißt, die Chance (das Risiko) als Fünfzigjähriger diese Krankheit zu bekommen beträgt nur 6,25% der Chance als Siebzigjähriger daran zu erkranken.

Die Varianz-Kovarianzmatrix

Es seien $x_1, x_2, ..., x_n$ die Stichprobenwerte einer Zufallsvariablen X. Die Varianz σ^2 berechnet sich dann zu

$$\boxed{\mathrm{Var}(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 = \mu_{X^2} - \mu_X^2 = \overline{x^2} - \bar{x}^2}.$$

Zusätzlich seien $y_1, y_2, ..., y_n$ die Stichprobenwerte einer Zufallsvariablen Y. Um den Zusammenhang zwischen X und Y zu charakterisieren, verwendet man die Kovarianz

$$\boxed{\mathrm{Cov}(X,Y) = \sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)\cdot(y_i - \mu_Y) = \frac{1}{n}\sum_{i=1}^{n}x_i \cdot y_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right) = \mu_{X\cdot Y} - \mu_X \cdot \mu_Y = \overline{x \cdot y} - \bar{x} \cdot \bar{y}}.$$

Daraus folgt $\boxed{\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)}$ und $\boxed{\mathrm{Cov}(X,X) = \mathrm{Var}(X)}$.

Beispiel: Abschlusstabelle 2019/2020 der 1. Bundesliga.

Pl.	Verein	Sp.	S	U	N	Tore	Diff.	Punkte
1.	FC Bayern München (M, P)	34	26	4	4	100:32	+68	82
2.	Borussia Dortmund	34	21	6	7	84:41	+43	69
4.	Borussia Mönchengladbach	34	20	5	9	66:40	+26	65
5.	Bayer 04 Leverkusen	34	19	6	9	61:44	+17	63
3.	RB Leipzig	34	18	12	4	81:37	+44	66
6.	TSG 1899 Hoffenheim	34	15	7	12	53:53	±0	52
7.	VfL Wolfsburg	34	13	10	11	48:46	+2	49
8.	SC Freiburg	34	13	9	12	48:47	+1	48
9.	Eintracht Frankfurt	34	13	6	15	59:60	−1	45
11.	1. FC Union Berlin (N)	34	12	5	17	41:58	−17	41
10.	Hertha BSC	34	11	8	15	48:59	−11	41
13.	1. FSV Mainz 05	34	11	4	19	44:65	−21	37
14.	1. FC Köln (N)	34	10	6	18	51:69	−18	36
12.	FC Schalke 04	34	9	12	13	38:58	−20	39
15.	FC Augsburg	34	9	9	16	45:63	−18	36
16.	Werder Bremen	34	8	7	19	42:69	−27	31
17.	Fortuna Düsseldorf	34	6	12	16	36:67	−31	30
18.	SC Paderborn 07 (N)	34	4	8	22	37:74	−37	20

Bedeutung der Spalten:

Sp.: Anzahl der Spiele
S: Anzahl der Siege
U: Anzahl der Unentschieden
N: Anzahl der Niederlagen
Tore: Verhältnis Tore : Gegentore
Diff.: Anzahl Tore – Anzahl Gegentore.

(M) bedeutet Meister
(P) bedeutet Pokalsieger
(N) bedeutet Aufsteiger.

Es soll nun untersucht werden, ob es annähernd lineare Zusammenhänge zwischen den drei Datensätzen der Tore, Gegentore und Punkte gibt.

Im Folgenden bedeuten die drei Zufallsvariablen T, G und P die Tore, Gegentore und Punkte. Es ergibt sich $\overline{T} = \frac{1}{18}\sum_{i=1}^{18} T_i = \frac{982}{18} \approx 54{,}56$, d.h. in der Saison 2019/2020 hat eine Mannschaft im Mittel 54.56 Tore erzielt.

Natürlich ist dann $\overline{G} = \frac{1}{18}\sum_{i=1}^{18} G_i = \frac{982}{18} \approx 54{,}56$ gleich groß. $\overline{P}$ ergibt sich zu $\overline{P} = \sum_{i=1}^{18} P_i = \frac{850}{18} \approx 47{,}22$ als mittlere Punktzahl pro Mannschaft.

Für die Varianz ergibt sich $\mathrm{Var}(T) = \frac{1}{17}\sum_{i=1}^{18}(T_i - \overline{T})^2 = 319{,}90850$. Der Grund für die Division durch 17 statt durch 18 wird im Kapitel „Induktive Statistik“ hergeleitet: Bei einer Division durch 18 erhält man die Varianz der 18 gegebenen Daten. Bei Division durch 17 vergrößert sich das Ergebnis etwas und man erhält auf diese Weise einen Schätzwert der Varianz für eine große Zahl von Abschlusstabellen.

Analog folgt $\mathrm{Var}(G) = \frac{1}{17}\sum_{i=1}^{18}(G_i - \overline{G})^2 = 156{,}49673$ und $\mathrm{Var}(P) = \frac{1}{17}\sum_{i=1}^{18}(P_i - \overline{P})^2 = 260{,}88889$.

Für die geschätzten Kovarianzen gilt $\mathrm{Cov}(T,G) = \frac{1}{17}\sum_{i=1}^{18}(T_i - \overline{T})\cdot(G_i - \overline{G}) = -183{,}50327$,

$\mathrm{Cov}(T,P) = \frac{1}{17}\sum_{i=1}^{18}(T_i - \overline{T})\cdot(P_i - \overline{P}) = 265{,}10458$ und $\mathrm{Cov}(G,P) = \frac{1}{17}\sum_{i=1}^{18}(G_i - \overline{G})\cdot(P_i - \overline{P}) = -193{,}77778$.

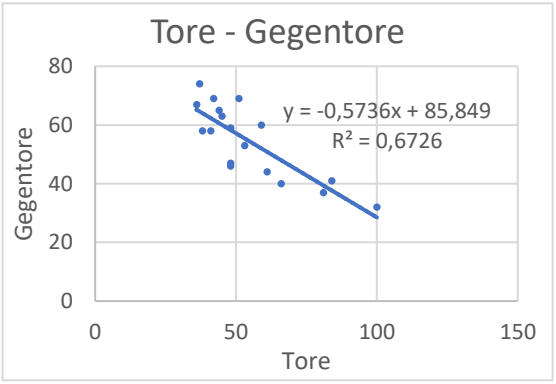
Nun lässt sich die Varianz-Kovarianzmatrix erstellen:

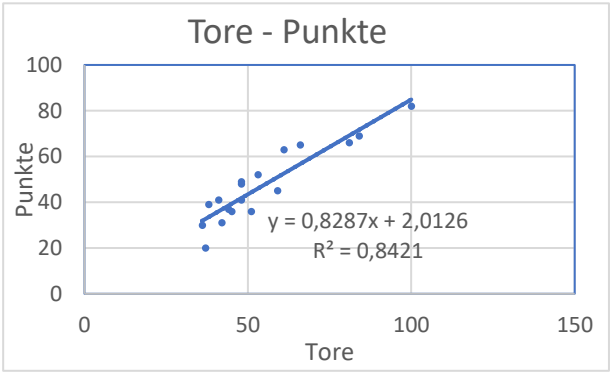
$$\sum = \begin{pmatrix} \mathrm{Var}(T) & \mathrm{Cov}(T,G) & \mathrm{Cov}(T,P) \\ \mathrm{Cov}(G,T) & \mathrm{Var}(G) & \mathrm{Cov}(G,P) \\ \mathrm{Cov}(P,T) & \mathrm{Cov}(P,G) & \mathrm{Var}(P) \end{pmatrix} = \begin{pmatrix} 319{,}90850 & -183{,}50327 & 265{,}10458 \\ -183{,}50327 & 156{,}49673 & -193{,}77778 \\ 265{,}10458 & -193{,}77778 & 260{,}88889 \end{pmatrix}.$$

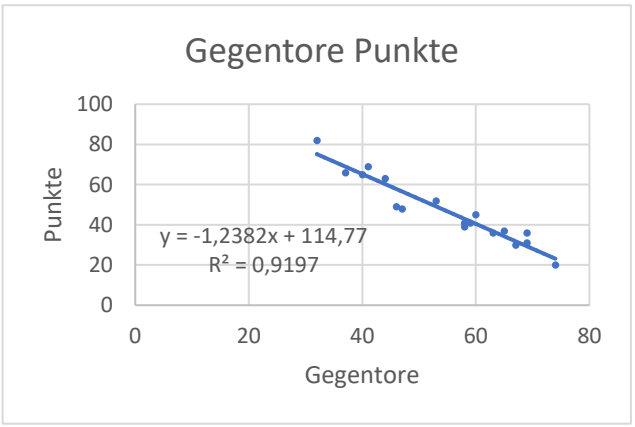
Noch aussagekräftiger für uns ist die Korrelationsmatrix R nach Bravais – Pearson mit den Korrelationskoeffizienten $r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$. Für diese Koeffizienten r gilt $-1 \le r \le 1$, und je näher r bei 1 bzw. -1 liegt, desto genauer liegen die betreffenden Punkte auf einer Geraden mit positiver bzw. negativer Steigung.

$$R = \begin{pmatrix} r_{T,T} & r_{T,G} & r_{T,P} \\ r_{G,T} & r_{G,G} & r_{G,P} \\ r_{P,T} & r_{P,G} & r_{P,P} \end{pmatrix} = \begin{pmatrix} 1 & -0,82012165 & 0,91764889 \\ -0,82012165 & 1 & -0,9590109 \\ 0,91764889 & -0,9590109 & 1 \end{pmatrix}.$$

Excel gibt nicht den Korrelationskoeffizienten $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$ an, sondern sein Quadrat $R^2 = r_{XY}^2$, das sog. „Bestimmtheitsmaß“ an.







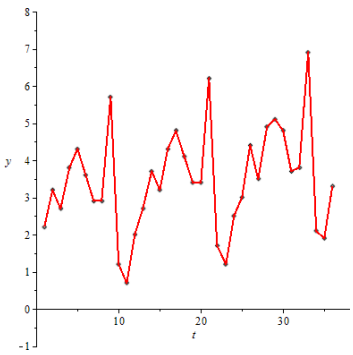
Die Zeitreihenanalyse

Betrachtet wird ein Merkmal über einen längeren Zeitraum in regelmäßigen Zeitabständen. Daher der Name „Zeitreihe“. Diese Reihe soll nun auf Gesetzmäßigkeiten untersucht (analysiert) werden, um eventuell Aussagen über den weiteren Verlauf zu machen.

Beispiel: Der Umsatz $y(t)$ einer Firma wird über drei Jahre monatlich, $t = 1..36$ festgestellt.

t	1	2	3	4	5	6	7	8	9	10	11	12
y(t)	**2,2**	**3,2**	**2,7**	**3,8**	**4,3**	**3,6**	**2,9**	**2,9**	**5,7**	**1,2**	**0,7**	**2,0**
t	13	14	15	16	17	18	19	20	21	22	23	24
y(t)	**2,7**	**3,7**	**3,2**	**4,3**	**4,8**	**4,1**	**3,4**	**3,4**	**6,2**	**1,7**	**1,2**	**2,5**
t	25	26	27	28	29	30	31	32	33	34	35	36
y(t)	**3,0**	**4,4**	**3,5**	**4,9**	**5,1**	**4,8**	**3,7**	**3,8**	**6,9**	**2,1**	**1,9**	**3,3**

Zur Analyse steht das **additive Zeitreihenmodell** $\boxed{y(t) = g(t) + s(t) + r(t)}$ zur Verfügung.



Dabei ist $g(t)$ die **„glatte Komponente“**. Sie kann z.B. linear $g(t) = a + b \cdot t$, quadratisch $g(t) = a + b \cdot t + c \cdot t^2$ oder auch exponentiell $g(t) = a \cdot e^{b \cdot t}$ sein.

$s(t)$ ist die **„saisonale Komponente“**.

$r(t)$ sind die **„Reste“**. Sie sind ein Maß für die Güte des Modells.

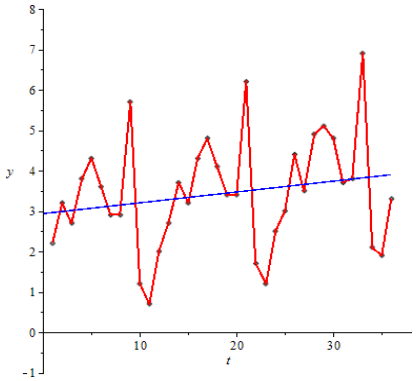
Daneben gibt es noch das multiplikative Zeitreihenmodell $y(t) = g(t) \cdot s(t) \cdot r(t)$.
Durch Logarithmieren wird es auf das additive Modell zurückgeführt.

Wir verwenden hier das additive Modell mit der linearen Komponente $g(t)$. Diese erhalten wir durch die lineare Regression mit den Formeln $b = \frac{s_{XY}}{s_X^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$ und $a = \bar{y} - b \cdot \bar{x}$. Nur verwenden wir t statt x.

Es folgt $\overline{t \cdot y} = \frac{1}{36} \sum_{i=1}^{36} t_i \cdot y_i = \frac{2394,1}{36} = \frac{23941}{360} \approx 66,503$.

$\bar{t} = \frac{1}{36} \sum_{i=1}^{36} t_i = \frac{666}{36} = \frac{37}{2} = 18,5$, $\bar{y} = \frac{1}{36} \sum_{i=1}^{36} y_i = \frac{123,8}{36} = \frac{619}{180} \approx 3,439$,

$\overline{t^2} = \frac{1}{36} \sum_{i=1}^{36} t_i^2 = \frac{16206}{36} = \frac{2701}{6} \approx 450,17$.

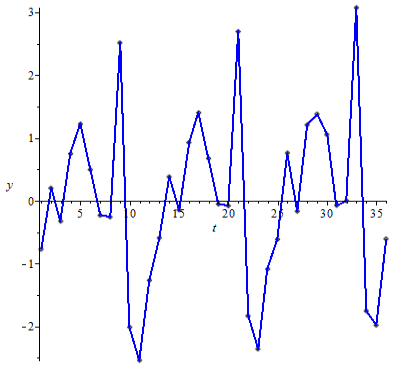


Und damit $b = \frac{173/60}{1295/12} = \frac{173}{6475} \approx 0,026718$ und

$a = \frac{619}{180} - \frac{173}{6475} \cdot \frac{37}{2} = \frac{18551}{6300} \approx 2,9446$,

so dass $g(t) = 2,9446 + 0,026718 \cdot t$.

Das Schaubild zeigt die Funktion $y^*(t) = y(t) - g(t) = s(t) + r(t)$.



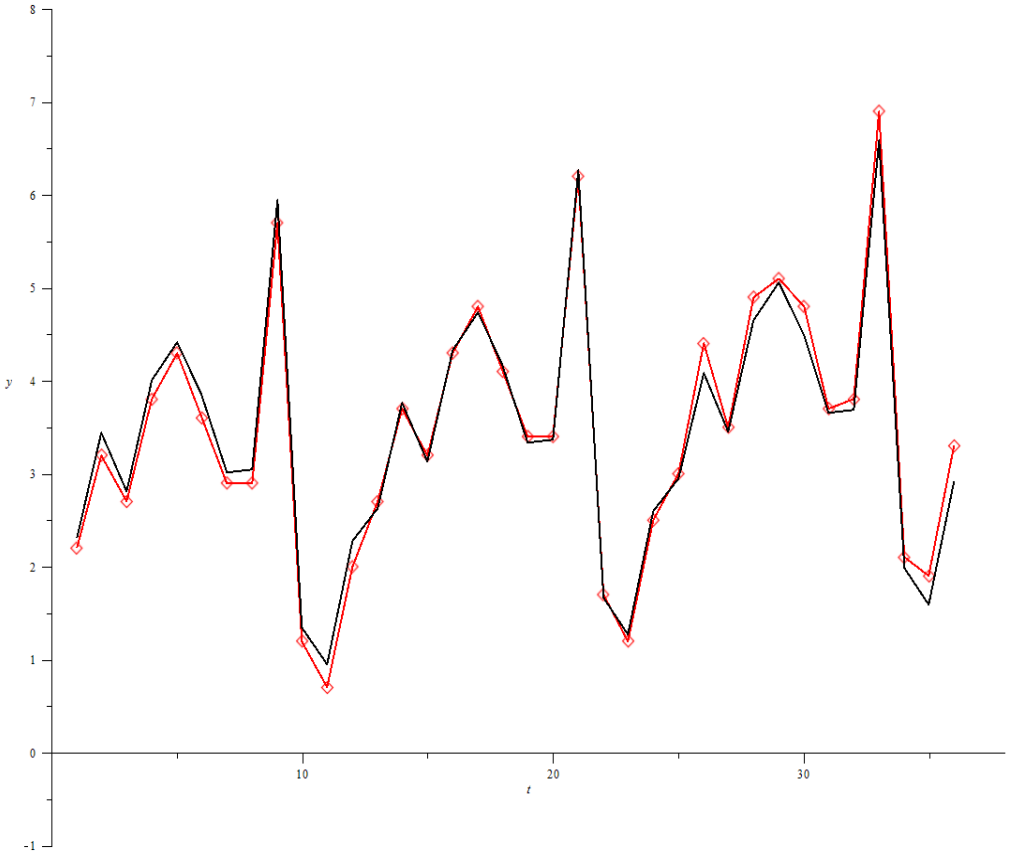
Wie erhält man daraus die saisonale Komponente?

Wir nehmen an, dass diese Komponente die Periode 1 Jahr besitzt, dass also $s(1) = s(13) = s(25)$, $s(2) = s(14) = s(26)$, $s(3) = s(15) = s(27)$, ..., $s(12) = s(24) = s(36)$ gilt.

Deshalb wählt man $s(1) = s(13) = s(25) = \frac{1}{3}\left(y^*(1) + y^*(13) + y^*(25)\right)$,

$s(2) = s(14) = s(26) = \frac{1}{3}\left(y^*(2) + y^*(14) + y^*(26)\right)$, usw.

t	1	2	3	4	5	6	7	8	9	10	11	12
y(t)	**2,2**	**3,2**	**2,7**	**3,8**	**4,3**	**3,6**	**2,9**	**2,9**	**5,7**	**1,2**	**0,7**	**2,0**
g(t)	2,971	2,998	3,025	3,051	3,078	3,105	3,132	1,158	3,185	3,212	3,239	3,265
y*(t)	–0,771	0,202	–0,325	0,749	1,222	0,495	–0,232	–0,258	2,515	–2,012	–2,539	–1,265
t	13	14	15	16	17	18	19	20	21	22	23	24
y(t)	**2,7**	**3,7**	**3,2**	**4,3**	**4,8**	**4,1**	**3,4**	**3,4**	**6,2**	**1,7**	**1,2**	**2,5**
g(t)	3,292	3,319	3,345	3,372	3,399	3,426	3,452	3,479	3,506	3,532	3,559	3,586
y*(t)	–0,592	0,381	–0,145	0,928	1,401	0,674	–0,052	–0,079	2,694	–1,832	–2,359	–1,086
t	25	26	27	28	29	30	31	32	33	34	35	36
y(t)	**3,0**	**4,4**	**3,5**	**4,9**	**5,1**	**4,8**	**3,7**	**3,8**	**6,9**	**2,1**	**1,9**	**3,3**
g(t)	3,613	3,639	3,666	3,693	3,719	3,746	3,773	3,800	3,826	3,853	3,880	3,906
y*(t)	–0,613	0,761	–0,166	1,207	1,381	1,054	–0,073	0,000	3,074	–1,753	–1,980	–0,606
s(t)	–0,659	0,448	–0,212	0,961	1,335	0,741	–0,119	–0,112	2,761	–1,866	–2,292	–0,986



Die roten Punkte waren gegeben, der rote Streckenzug verbindet sie. Der schwarze Streckenzug gehört zu $y = g(t) + s(t)$. Die Übereinstimmung ist verblüffend, der Rest $r(t)$ minimal.

Da $g(t)$ und $s(t)$ bekannt sind, kann man den schwarzen Streckenzug nach rechts extrapolieren.

Zum Beispiel: $y(37) = g(37) + s(37) = g(37) + s(1) = 3,933 - 0,659 = 3,274$

oder $y(38) = g(38) + s(38) = g(38) + s(2) = 3,960 + 0,448 = 4,408$.