1. Das Statistische Landesamt Baden-Württemberg veröffentlichte die Zahlen für die Übergänge aus Klassenstufe 4 an den Grundschulen in Baden-Württemberg zum Schuljahr 2008/09 nach Nationalität unterschieden. Die angegebenen Zahlen habe ich auf Hunderter gerundet.

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	20200	33300	40500	800	94800
Ausländer	7300	3800	2800	200	14100
$\sum$	27500	37100	43300	1000	108900

Bestimmen Sie die bedingten Verteilungen für die Zeilen.

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	20200/94800 = 0,213	33300/94800 = 0,351	40500/94800 = 0,427	800/94800 = 0,008	1
Ausländer	7300/14100 = 0,518	3800/14100 = 0,270	2800/14100 = 0,199	200/14100 = 0,014	1
$\sum$	27500/108900 = 0,253	37100/108900 = 0,341	43300/108900 = 0,398	1000/108900 = 0,009	1

Die beiden Werte pro Spalte unterscheiden sich sehr deutlich. Unabhängigkeit ist nicht gegeben.

Bestimmen Sie die bedingten Verteilungen für die Spalten.

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	20200/27500 = 0,735	33300/37100 = 0,898	40500/43300 = 0,935	800/1000 = 0,800	94800/108900 = 0,871
Ausländer	7300/27500 = 0,265	3800/37100 = 0,102	2800/43300 = 0,065	200/1000 = 0,200	14100/108900 = 0,129
$\sum$	1	1	1	1	1

Die vier Werte pro Zeile unterscheiden sich ebenfalls sehr deutlich. Unabhängigkeit ist nicht gegeben.

Vervollständigen Sie die Tabelle so, dass die beiden Variablen Schulart und Herkunft unabhängig sind.

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	23939	32296	37694	871	94800
Ausländer	3561	4804	5606	129	14100
$\sum$	27500	37100	43300	1000	108900

Zur Übung wird jetzt die Unabhängigkeit explizit nachgerechnet:

Bedingte Verteilungen für die Zeilen:

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	23939/94800 = 0,253	32296/94800 = 0,341	37694/94800 = 0,398	871/94800 = 0,009	1
Ausländer	3561/14100 = 0,253	4804/14100 = 0,341	5606/14100 = 0,398	129/14100 = 0,009	1
$\sum$	27500/108900 = 0,253	37100/108900 = 0,341	43300/108900 = 0,398	1000/108900 = 0,009	1

Und für die bedingten Verteilungen für die Spalten:

	Hauptschulen	Realschulen	Gymnasien	Sonstige	$\sum$
Deutsche	23939/27500 = 0,871	32296/37100 = 0,871	37694/43300 = 0,871	871/1000 = 0,871	94800/108900 = 0,871
Ausländer	3561/27500 = 0,129	4804/37100 = 0,129	5606/43300 = 0,129	129/1000 = 0,129	14100/108900 = 0,129
$\sum$	1	1	1	1	1

Bestimmen Sie entsprechend der Tabelle im Skript Seite 4 den χ^2 – Koeffizienten und den korrigierten Kontingenzkoeffizienten C^*.

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

	Hauptschulen		Realschulen		Gymnasien		Sonstige		$\sum$
Deutsche	20200	23939	33300	32296	40500	37694	800	871	94800
	13980121	584	1008016	31	7873636	209	5041	6	
Ausländer	7300	3561	3800	4804	2800	5606	200	129	14100
	13980121	3926	1008016	210	7873636	1405	5041	39	
$\sum$	27500		37100		43300		1000		108900

$\chi^2 = (584+31+209+6)+(3926+210+1405+39) = 6410$.

Der Kontingenzkoeffizient wird $C = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{6410}{6410+108900}} \approx 0,236$ und der korrigierte C-Koeffizient $C^* = \sqrt{\frac{M}{M-1}\cdot\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{2}{1}\cdot\frac{6410}{6410+108900}} \approx 0,333$ ist deutlich größer als Null, was die Abhängigkeit der beiden Variablen beweist.

2. Es wurden $n = 100$ Teile bezüglich der Eigenschaften x und y untersucht, siehe Tabelle.

	x_1		x_2		x_3		
y_1	1	2	6	6	13	12	20
	1	1/2	0	0	1	1/12	
y_2	5	3	10	9	15	18	30
	4	4/3	1	1/9	9	1/2	
y_3	4	5	14	15	32	30	50
	1	1/5	1	1/15	4	2/15	
	10		30		60		$n = 100$

Wieviel Prozent aller Teile haben die Eigenschaft y_2 ? 30%
Wieviel Prozent aller Teile haben die Eigenschaften x_3 und y_2 ? 15%
Wieviel Prozent der Teile mit der Eigenschaft y_3 haben die Eigenschaft x_2 ? 28%
Wieviel Prozent der Teile mit der Eigenschaft x_1 haben die Eigenschaft y_3 ? 40%

$\chi^2 = (1/2+0+1/12)+(4/3+1/9+1/2)+(1/5+1/15+2/15) = 7/12+35/18+2/5 = 527/180$.

$C^* = \sqrt{\frac{M}{M-1}\cdot\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{3}{2}\cdot\frac{527/180}{527/180+100}} = \sqrt{\frac{1581}{37054}} \approx 0,207$, d.h. leichte Abhängigkeit der beiden Eigenschaften.

3. Bestimmen Sie die Regressionsgerade $y = a + b \cdot x$ durch die vier gegebenen Punkte. Bestimmen Sie den Korrelationskoeffizienten r.

x_i	-2	0	1	3
y_i	8	2	-1	-7

Es ist $\overline{x\cdot y} = \frac{1}{4}(-16+0-1-21) = -\frac{19}{2}$

$\bar{x} = \frac{1}{4}\cdot 2 = \frac{1}{2}$, $\bar{y} = \frac{1}{4}\cdot 2 = \frac{1}{2}$, $\overline{x^2} = \frac{1}{4}\cdot(4+0+1+9) = \frac{7}{2}$, $\overline{y^2} = \frac{1}{4}\cdot(64+4+1+49) = \frac{59}{2}$. . Damit folgt

$$b = \frac{\overline{x\cdot y} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{-\frac{19}{2} - \frac{1}{2}\cdot\frac{1}{2}}{\frac{7}{2} - \frac{1}{4}} = \frac{-\frac{39}{4}}{\frac{13}{4}} = -3 \quad \text{und} \quad a = \bar{y} - b\cdot\bar{x} = \frac{1}{2} - (-3)\cdot\frac{1}{2} = 2\,.$$

Folglich lautet die Gleichung der Regressionsgerade $y = 2 - 3x$.

Korrelationskoeffizient $r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\overline{x\cdot y} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{-\frac{19}{2} - \frac{1}{2}\cdot\frac{1}{2}}{\sqrt{\frac{7}{2} - \frac{1}{4}}\cdot\sqrt{\frac{59}{2} - \frac{1}{4}}} = \frac{-\frac{39}{4}}{\sqrt{\frac{13}{4}\cdot\frac{117}{4}}} = \frac{-\frac{39}{4}}{\frac{39}{4}} = -1\,.$

Die vier Punkte liegen sogar exakt auf der absteigenden Geraden $y = 2 - 3x$.

4. Die drei Punkte $(1/4)$, $(2/4)$, $(3/4)$ liegen auf der horizontalen Geraden $y = 4$. Bestimmen Sie zur Probe die Gleichung $y = a + b \cdot x$ der Regressionsgeraden und den Korrelationskoeffizienten r.

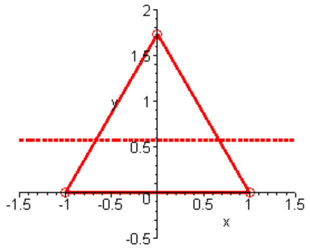
$\overline{x\cdot y} = \frac{1}{3}\cdot(4+8+12) = 8$, $\bar{x} = \frac{1}{3}\cdot(1+2+3) = 2$, $\bar{y} = \frac{1}{3}\cdot(4+4+4) = 4$, sdff $\overline{x^2} = \frac{1}{3}\cdot(1+4+9) = \frac{14}{3}$.

$\overline{y^2} = \frac{1}{3}\cdot(16+16+16) = 16$. Folglich $b = \frac{s_{XY}}{s_X^2} = \frac{\overline{x\cdot y} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{8 - 2\cdot 4}{\frac{14}{3} - 4} = 0$, $a = \bar{y} - b\cdot\bar{x} = 4 - 0\cdot 2 = 4$.

Daraus folgt die Regressionsgerade zu $y = 4$ und der Korrelationskoeffizient r zu

$r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\overline{x\cdot y} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}} = \frac{8 - 2\cdot 4}{\sqrt{\frac{2}{3}}\cdot\sqrt{16-16}} = \frac{0}{0}$ und nicht definiert.

5. Die drei Punkte $(1/4)$, $(2/4)$, $(3/4)$ liegen auf der horizontalen Geraden $y = 4$. Bestimmen Sie zur Probe die Gleichung $y = a + b \cdot x$ der Regressionsgeraden und den Korrelationskoeffizienten r.



i	x_i	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	-1	0	1	0	0		-1	1	$\frac{1}{3}\sqrt{3}$
2	1	0	1	0	0		1	1	$-\frac{1}{3}\sqrt{3}$
3	0	$\sqrt{3}$	0	3	0		0	0	0
$\sum$	**0**	$\boldsymbol{\sqrt{3}}$	**2**	**3**	**0**		**0**	**2**	**0**

Es gilt $\bar{x} = \frac{1}{3}(-1+1+0) = 0$, $\bar{y} = \frac{1}{3}\cdot(0+0+\sqrt{3}) = \frac{1}{3}\sqrt{3}$, $\sum_{i=1}^{3} x_i \cdot y_i = (-1)\cdot 0 + 1\cdot 0 + 0\cdot\sqrt{3} = 0$,

also $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x}\cdot\bar{y} = 0$.

Außerdem $\sum_{i=1}^{3} x_i^2 = (-1)^2 + 0^2 + 1^2 = 2$, so dass $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \overline{x^2} - \bar{x}^2 = \frac{2}{3} - 0 = \frac{2}{3}$ wird.

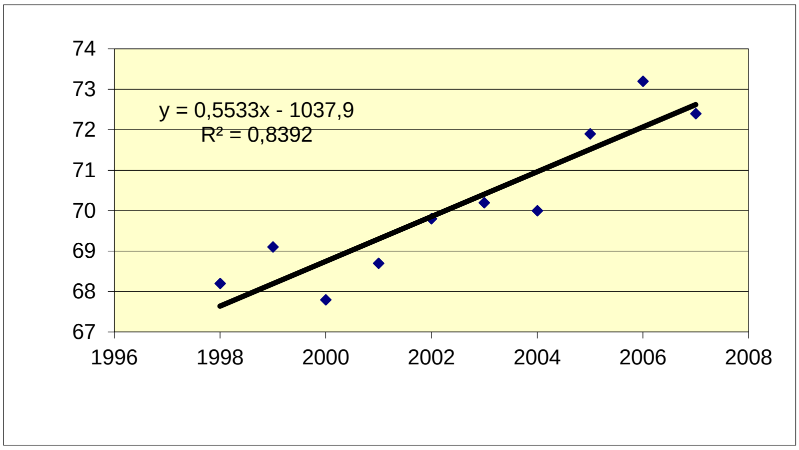
Es folgt $b = \frac{s_{XY}}{s_X^2} = 0$ und $a = \bar{y} - b\cdot\bar{x} = \frac{1}{3}\sqrt{3}$, so dass die Regressionsgerade $y = \frac{1}{3}\sqrt{3}$ lautet.

Wegen $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \overline{y^2} - \bar{y}^2 = \frac{1}{3}\cdot(0^2+0^2+\sqrt{3}^2) - \left(\frac{1}{3}\sqrt{3}\right)^2 = \frac{2}{3}$ folgt der Korrelationskoeffizient

zu $r = \frac{s_{XY}}{s_X \cdot s_y} = \frac{0}{\sqrt{2/3}\cdot\sqrt{2/3}} = 0$ wird. Und damit besteht erwartungsgemäß keine Korrelation zwischen den drei Punkten.

6. Die Tabelle gibt die Bruttostromerzeugung in Baden-Württemberg in Milliarden kWh für die angegebenen Jahre an.

Jahr x_i	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Strom y_i	68,2	69,1	67,8	68,7	69,8	70,2	70,0	71,9	73,2	72,4



Excel liefert die Gleichung $y = 0{,}5533\cdot x - 1037{,}9$ und das Bestimmtheitsmaß $R^2 = 0{,}8392$.

Bestimmen Sie die Gleichung der Regressionsgeraden. Erstellen Sie dazu eine Tabelle wie im Skript Seite 5.

i	Jahr x_i	Strom 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	68,2	3992004	4651,24	136263,6		-4,5	20,25	8,685
2	1999	69,1	3996001	4774,81	138130,9		-3,5	12,25	3,605
3	2000	67,8	4000000	4596,84	135600,0		-2,5	6,25	5,825
4	2001	68,7	4004001	4719,69	137468,7		-1,5	2,25	2,145
5	2002	69,8	4008004	4872,04	139739,6		-0,5	0,25	0,165
6	2003	70,2	4012009	4928,04	140610,6		0,5	0,25	0,035
7	2004	70,0	4016016	4900,00	140280,0		1,5	2,25	-0,195
8	2005	71,9	4020025	5169,61	144159,5		2,5	6,25	4,425
9	2006	73,2	4024036	5358,24	146839,2		3,5	12,25	10,745
10	2007	72,4	4028049	5241,76	145306,8		4,5	20,25	10,215
$\sum$	**20025**	**701,3**	**40100145**	**49212,27**	**1404398,9**		**0**	82,5	**45,65**

Es ist $\bar{x} = 20025/10 = 2002{,}5$ und $\bar{y} = 701{,}3/10 = 70{,}13$.

$s_{XY} = \frac{1}{n} \cdot \sum_{i=1}^{10} x_i \cdot y_i - \bar{x} \cdot \bar{y} = \frac{1}{10} \cdot 1404398{,}9 - 2002{,}5 \cdot 70{,}13 = 4{,}565$.

$s_X^2 = \frac{1}{10} \cdot \sum_{i=1}^{10} x_i^2 - \bar{x}^2 = \frac{1}{10} \cdot 40100145 - 2002{,}5^2 = 8{,}25$.

Und damit $b = \frac{s_{XY}}{s_X^2} = \frac{4{,}565}{8{,}25} = \frac{83}{150} \approx 0{,}5533$ und $a = \bar{y} - b \cdot \bar{x} = 70{,}13 - \frac{83}{150} \cdot 2002{,}5 = -1037{,}92$.

Und damit lautet die Geradengleichung $y = \frac{83}{150} \cdot x - 1037{,}92$.

Bestimmen Sie den Korrelationskoeffizienten r.

Wegen $s_Y^2 = \frac{1}{10} \cdot \sum_{i=1}^{10} y_i^2 - \bar{y}^2 = \frac{1}{10} \cdot 49212{,}27 - 70{,}13^2 = 3{,}0101$ folgt der Korrelationskoeffizient zu

$r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{4{,}565}{\sqrt{8{,}25 \cdot 3{,}0101}} = 0{,}9161$ bzw. $r^2 = 0{,}8392$, was Excel angibt.

Da r in der Nähe von 1 liegt, besteht ein guter ansteigender linearer Zusammenhang zwischen der Stromerzeugung und der Jahreszahl.

Mit welcher Stromerzeugung ist im Jahre 2010 zu rechnen?

$y = \frac{83}{150} \cdot 2010 - 1037{,}92 = 74{,}28$ Milliarden kWh.

In welchem Jahr müsste die Stromerzeugung bei 75,0 Milliarden kWh liegen?

Aus $\frac{83}{150} \cdot x - 1037{,}92 = 75$ ergibt sich $x = 2011{,}3$.

7. Bei einer Schulklasse wurde der Zusammenhang zwischen dem täglichen Fernsehkonsum X in Minuten und dem Körpergewicht Y in kg untersucht. Es ergab sich $s_X^2 = 1369$, $s_Y^2 = 36$ und $s_{XY} = 180$. Was lässt sich über den Grad der linearen Abhängigkeit von X und Y aussagen?

   Der Korrelationskoeffizient beträgt $r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{180}{\sqrt{1396 \cdot 36}} = \frac{180}{37 \cdot 6} \approx 0{,}81$, d.h. ein recht deutlicher linearer ansteigender Zusammenhang zwischen X und Y.

8. Es werden n Datenpunkte $P_i(x_i, y_i)$, $x_i > 0$, $n \in \mathbb{N}$, erhoben. Sie sollten einer Gleichung der Form $y = a \cdot \ln x$, $a \in \mathbb{R}$, genügen. Bestimmen Sie mit Hilfe der Methode der kleinsten Quadrate diejenige Formel, mit der sich die Konstante a bestimmen lässt.

Durch geeignete Wahl von a muss die Summe $\sum_{i=1}^{n} (y_i - a \cdot \ln x_i)^2$ minimal werden.

Aus $\frac{d}{da} \sum_{i=1}^{n} (y_i - a \cdot \ln x_i)^2 = \sum_{i=1}^{n} 2 \cdot (y_i - a \cdot \ln x_i) \cdot (-\ln x_i) = 0$ folgt

$\sum_{i=1}^{n} (y_i - a \cdot \ln x_i) \cdot (-\ln x_i) = -\sum_{i=1}^{n} (y_i \cdot \ln x_i) + a \cdot \sum_{i=1}^{n} (\ln x_i)^2 = 0$, so dass $a = \frac{\sum_{i=1}^{n} (y_i \cdot \ln x_i)}{\sum_{i=1}^{n} (\ln x_i)^2}$.

Wegen $\frac{d^2}{da^2} \sum_{i=1}^{n} (y_i - a \cdot \ln x_i)^2 = \frac{d}{da} \sum_{i=1}^{n} 2 \cdot (y_i - a \cdot \ln x_i) \cdot (-\ln x_i) = \sum_{i=1}^{n} 2 \cdot (-\ln x_i)^2 > 0$ liegt wirklich ein Minimum vor.

9. Es sind die vier Datenpunkte gegeben:

x	-4	-1	5	8
y	$-1/2$	-1	1	$1/2$

Sie sollen einer Gleichung der Form $y = \frac{a}{b+x}$ mit $a, b \in \mathbb{R}$ genügen. Bestimmen Sie mit Hilfe der Methode der kleinsten Quadrate die Konstanten a und b.

Direkter Weg: Durch geeignete Wahl von a und b muss die Summe $\sum_{i=1}^{n}\left(y_i - \frac{a}{b+x_i}\right)^2$ minimal werden.

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

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}\left(y_i - \frac{a}{b+x_i}\right)^2 = 2a\sum_{i=1}^{n}\left(y_i - \frac{a}{b+x_i}\right)\cdot\frac{1}{(b+x_i)^2} = 0\text{, d.h. } \sum_{i=1}^{n}\frac{y_i}{(b+x_i)^2} = a\sum_{i=1}^{n}\frac{1}{(b+x_i)^3}\ .$$

Diese Formeln lassen sich nicht nach b auflösen.

x	-4	-1	5	8
y	$-1/2$	-1	1	$1/2$
z	-2	-1	1	2

Dabei gilt $z = \frac{1}{y} = \frac{b+x}{a} = \frac{b}{a} + \frac{1}{a}\cdot x$.

Nach unseren Formeln der linearen Regression gilt

$$\frac{1}{a} = \frac{s_{XZ}}{s_X^2} = \ = \frac{\overline{x\cdot z} - \overline{x}\cdot\overline{z}}{\overline{x^2} - \overline{x}^2} = \frac{\frac{15}{2} - 2\cdot 0}{\frac{53}{2} - 4} = \frac{1}{3}\text{, so dass } a = 3 \text{ folgt.}$$

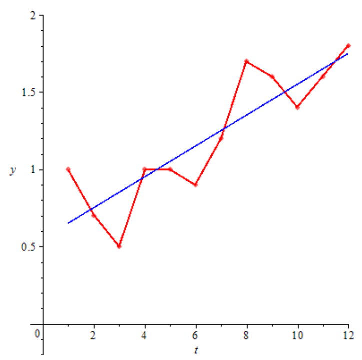
$$\frac{b}{a} = \overline{z} - \frac{1}{a}\cdot\overline{x} = 0 - \frac{1}{3}\cdot 2 = -\frac{2}{3}\text{, so dass } b = -2\ .$$

In der Tat, alle Punkte liegen exakt auf $y = \frac{a}{b+x} = \frac{3}{-2+x} = \frac{3}{x-2}$.

10. Der Umsatz $y(t)$ einer Firma wird quartalsweise über drei Jahre bestimmt. Analysieren Sie die gegebene Zeitreihe. Bestimmen Sie damit $y(t)$ für das 1. und 2. Quartal im 4. Jahr.

Jahr	Quartal			
	1	2	3	4
1	1,0	0,7	0,5	1,0
2	1,0	0,9	1,2	1,7
3	1,6	1,4	1,6	1,8

t	1	2	3	4
y(t)	**1,0**	**0,7**	**0,5**	**1,0**
y*(t)	0,35	–0,05	–0,35	0,05
t	5	6	7	8
y(t)	**1,0**	**0,9**	**1,2**	**1,7**
y*(t)	–0,05	–0,25	–0,05	0,35
t	9	10	11	12
y(t)	**1,6**	**1,4**	**1,6**	**1,8**
y*(t)	0,15	–0,15	–0,05	–0,05
s(t)	0,15	–0,15	–0,15	0,15



Bestimmung der Regressionsgeraden $g(t) = a + b \cdot t$: Es ist $b = \frac{s_{ty}}{s_t^2} = \frac{\overline{t \cdot y} - \bar{t} \cdot \bar{y}}{\overline{t^2} - \bar{t}^2} = \frac{\frac{1}{12} \cdot 107{,}9 - \frac{13}{2} \cdot \frac{6}{5}}{\frac{325}{6} - \frac{169}{4}} = \frac{1}{10}$ und

$a = \bar{y} - b \cdot \bar{t} = \frac{6}{5} - \frac{1}{10} \cdot \frac{13}{2} = \frac{11}{20}$, d.h. $g(t) = \frac{11}{20} + \frac{1}{10} t = 0{,}55 + 0{,}1 \cdot t$.

Es ist $y^*(t) = y(t) - g(t) = s(t) + r(t)$.

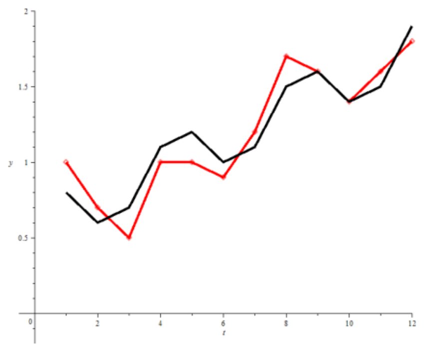
Bestimmung der saisonalen Komponente:
Wir nehmen an, dass diese Komponente die Periode 1 Jahr besitzt, also dass

$s(1) = s(5) = s(9) = \frac{1}{3}\left(y^*(1) + y^*(5) + y^*(9)\right) = 0{,}15$,

$s(2) = s(6) = s(10) = \frac{1}{3}\left(y^*(2) + y^*(6) + y^*(10)\right) = -0{,}15$,

$s(3) = s(7) = s(11) = \frac{1}{3}\left(y^*(3) + y^*(7) + y^*(11)\right) = -0{,}15$ und

$s(4) = s(8) = s(12) = \frac{1}{3}\left(y^*(4) + y^*(8) + y^*(12)\right) = 0{,}15$.



Im Schaubild sind $y(t)$ in rot mit den Punkten und die Näherung $g(t) + s(t)$ in schwarz ohne Punkte dargestellt.

Es ist $g(13) + s(13) = g(13) + s(1) = 1{,}85 + 0{,}15 = 2{,}00$,
$g(14) + s(14) = g(14) + s(2) = 1{,}95 - 0{,}15 = 1{,}80$.