Kapitel VI

Korrelations- und Regressionsanalyse

B. 6. 1. (*Gegenstand der Korrelations- und Regressionsanalyse*)
Während die *Korrelationsanalyse* die Existenz, die Stärke und die Richtung des Zusammenhangs zwischen zwei oder mehreren statistischen Variablen untersucht, ist Gegenstand der *Regressionsanalyse* die Darstellung des Zusammenhangs in funktionaler Form.

D. 6. 1. (*Bravais-Pearson-Korrelationskoeffizient*)
Die Merkmale X und Y seien kardinalskaliert. Als *Bravais-Pearson-Korrelationskoeffizient* bzw. *Produkt-Moment-Korrelationskoeffizient* bezeichnet man

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

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

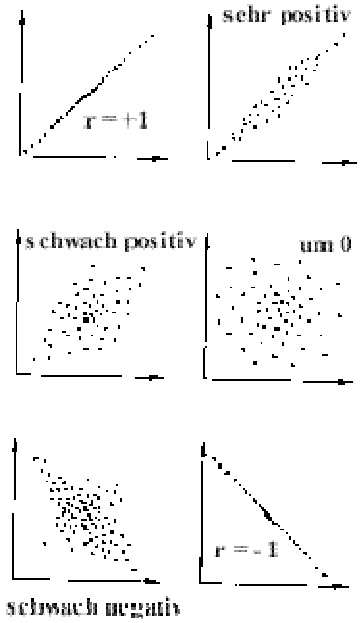
Dabei wird vorausgesetzt, dass weder alle $x_i -$ Werte noch alle $y_i -$ Werte gleich sind.

B. 6. 2.
1.
Es gilt

$$-1 \le r \le +1.$$

2.
$r < 0$ bedeutet einen gegenläufigen Zusammenhang; $r > 0$ einen gleichsinnigen.

3.
$r = 0$ zeigt keinen Zusammenhang und $r = 1$ einen ausschließlichen Zusammenhang.



BS. 6. 1.
Die nachfolgende Tabelle zeigt die Kurse zweier Aktien X und Y an 9 aufeinander folgenden Börsentagen:

Tag	1	2	3	4	5	6	7	8	9
Aktie X	5	6	11	8	13	8	10	16	13
Aktie Y	8	7	9	10	11	10	11	12	12

Berechnen und interpretieren Sie hierzu den entsprechenden Bravais-Pearson-Koeffizienten.

Lösung:

Arbeitstabelle

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
5	8	-5	-2	10	25	4
6	7	-4	-3	12	16	9
11	9	1	-1	-1	1	1
8	10	-2	0	0	4	0
13	11	3	1	3	9	1
8	10	-2	0	0	4	0
10	11	0	1	0	0	1
16	12	6	2	12	36	4
13	12	3	2	6	9	4
90	90	0	0	42	104	24

$$\bar{x}=\frac{90}{9}=10,\qquad \bar{y}=\frac{90}{9}=10,$$

$$r=\frac{\frac{42}{9}}{\sqrt{\frac{104}{9}\cdot\frac{24}{9}}}=0.841273288\approx 0.84.$$

Es handelt sich um einen starken gleichsinnigen Zusammenhang.

D. 6. 2. (*Rangkorrelationskoeffizient von Spearman*)
Die Merkmale X und Y seien ordinalskaliert. *Spearman-Rangkorrelationskoeffizient*. bezeichnet man

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - R_i')^2}{(n-1)n(n+1)}.$$

Dabei sind $R_i, R_i',\ i = 1,2,...,n,$ gewisse Rangnummern.

B. 6. 3.
Es gilt

$$-1 \le r_s \le +1.$$

2.
$r = 1$ bedeutet ein gleichsinniges Verhalten; $r = -1$ bedeutet gegensinniges Verhalten.

BS. 6. 2.
Zwei unabhängige Gutachter G_1 und G_2 sollen für eine Bank die Bonität von sieben Unternehmen beurteilen. Die Unternehmen werden dabei anhand eines Punktschemas von 1 (sehr schlechte Bonität) bis 10 (sehr gute Bonität) eingestuft. Das Ergebnis der Beurteilung ist in folgender Tabelle angegeben:

Unternehmen	Punkte von G_1	Punkte von G_2
1	2	3
2	3	2
3	3	4
4	6	6
5	7	5
6	8	8
7	9	10

Berechnen und interpretieren Sie hierzu den entsprechenden Korrelationskoeffizienten.

Lösung:

Unternehmen	Punkte von G_1	R_i	Punkte von G_2	R_i'
1	2	7	3	6
2	3	$5.5^{(*)}$	2	7
3	3	$5.5^{(*)}$	4	5
4	6	4	6	3
5	7	3	5	4
6	8	2	8	2
7	9	1	10	1

(*Hinweis*: Die Punkzahl 3 kommt zweimal in den Positionen 5 und 6 vor. Daher ist die Rangzahl gleich $\frac{5+6}{2} = 5.5$.)

R_i	$R_i^{'}$	$(R_i - R_i^{'})^2$
7	6	1.00
$5.5^{(*)}$	7	2.25
$5.5^{(*)}$	5	0.25
4	3	1.00
3	4	1.00
2	2	0.00
1	1	0.00
		5.50

$$r_s = 1 - \frac{6 \cdot 5.5}{6 \cdot 7 \cdot 8} = 0.901785714 \approx 0.90 .$$

Es gibt also große Übereinstimmung in der Beurteilung der Unternehmen durch die Gutachter.

D. 6. 3. (*Regressionsfunktion*)
Gegeben sei die zweidimensionale Verteilung der metrisch messbaren Merkmale X und Y. Es sei Y statistisch abhängig von X. Eine Funktion $y^* = f(x)$, die die Tendenz der Abhängigkeit beschreibt, heißt *Regressionsfunktion*.

B. 6. 3.
Die Koeffizienten der Regressionsfunktion werden folgendermaßen berechnet:

$$S(.) = \sum_{i=1}^{n} (y_i - y_i^*)^2 \rightarrow Min!$$

(*Methode der kleinsten Quadratsummen*)

B. 6. 4.
Die *Normalgleichungen* zur Bestimmung der *linearen Regressionsfunktion*

$$y^* = a_0 + a_1 x$$

lauten:

$$\begin{aligned} n \cdot a_0 \quad &+ a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i &+ a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{aligned}$$

S. 6. 1.
Für eine lineare Regressionsfunktion gilt

$$\bar{y} = a_0 + a_1 \bar{x} .$$

B e w e i s:
Wir dividieren beide Seiten der 1. Normalgleichung durch n :

$$\frac{\sum_{i=1}^{n} y_i}{n} = a_0 + a_1 \cdot \frac{\sum_{i=1}^{n} x_i}{n},$$

d. h.

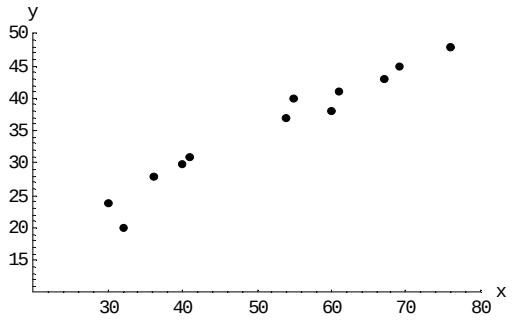
$$\bar{y} = a_0 + a_1 \bar{x} .$$

BS. 6. 3.
Es soll die Abhängigkeit des Niveaus der Arbeitsproduktivität von dem Automatisierungsgrad der Arbeit in 14 Betrieben untersucht werden. Dazu liegt folgendes Datenmaterial vor:

Betrieb	Niveau der Arbeitsproduktivität t/Std.	Automatisierungsgrad der Arbeit %
1	20	32
2	24	30
3	28	36
4	30	40
5	31	41
6	33	47
7	34	56
8	37	54
9	38	60
10	40	55
11	41	61
12	43	67
13	45	69
14	48	76

Lösung:

Das Streuungsdiagramm:



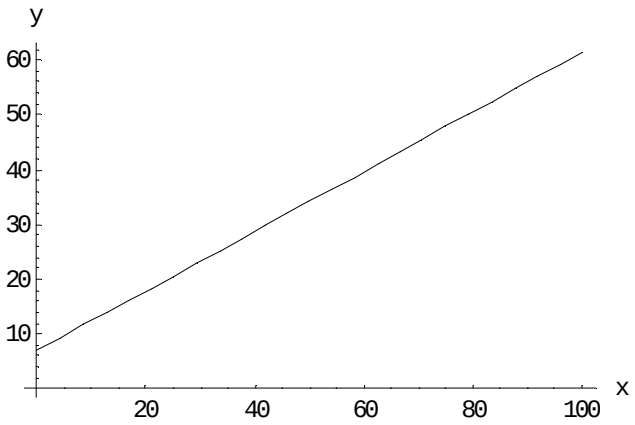
Arbeitstabelle

i	y_i	x_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	20	32	640	1024	400
2	26	30	720	900	576
3	28	36	1008	1296	784
4	30	40	1200	1600	900
5	31	41	1271	1681	961
6	33	47	1551	2209	1089
7	34	56	1904	3136	1156
8	37	54	1998	2916	1369
9	38	60	2280	3000	1444
10	40	55	2200	3025	1600
11	41	61	2501	3721	1681
12	43	67	2881	4489	1849
13	45	69	3105	4761	2025
14	48	76	3648	5776	2304
Summen	492	724	26907	40134	18138

$$\begin{aligned} 14a_0 \quad + 724a_1 &= 492 \\ 724a_0 + 40134a_1 &= 26907 \end{aligned} \quad \Rightarrow \quad a_0 = 7.0356, \quad a_1 = 0.5435\,.$$

Damit lautet die gesuchte lineare Regressionsfunktion:

$$y^* = 7.0356 + 0.5435x\,.$$



Beispielsweise gilt:

$y^*(42) = 7.0356 + 0.5435 \cdot 42 \approx 29.86$ (Interpolation)
$y^*(80) = 7.0356 + 0.5435 \cdot 80 \approx 50.52$ (Extrapolation).

<u>D. 6. 4.</u> (*Korrelationskoeffizient, Bestimmtheitsmaß nach Bravais-Pearson*)
1. Als *Korrelationskoeffizient zur linearen Regression* bezeichnet man

$$r_{yx} := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i\right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i\right)}}$$

$$-1 \le r_{yx} \le +1$$

2. Als *Bestimmtheitsmaß zur linearen Regression* bezeichnet man

$$B_{yx} := r_{yx}^2, \qquad 0 \le B_{yx} \le 1\,.$$

B. 6. 5.
Ein positiver Korrelationskoeffizient bedeutet gleichgerichtete Bewegung der Merkmale, ein negativer Korrelationskoeffizient entgegengesetzte Bewegung der Merkmale.
Das Bestimmtheitsmaß gibt das Gewicht des Merkmals X zur Bestimmung des Merkmals Y an.

BS. 6. 3. (*Fortsetzung***)**

$$r_{yx} := \frac{14 \cdot 26907 - 724 \cdot 492}{\sqrt{(14 \cdot 40134 - 724 \cdot 724) \cdot (14 \cdot 18138 - 492 \cdot 492)}} \approx 0.9687\,.$$

Der positive Korrelationskoeffizient besagt, dass mit zunehmendem Automatisierungsgrad die Arbeitsproduktivität auch zunimmt.

$$B_{yx} = 0.9384.$$

Die Entwicklung der Arbeitsproduktivität ist damit zu etwa 94% auf die Entwicklung des Automatisierungsgrad zurückzuführen.

BS. 6. 4.
Der pro- Kopf-Verbrauch an Butter (in kg) hat sich in einem Land in den Jahren 1999 – 2005 wie folgt entwickelt:

Jahr	1999	2000	2001	2002	2003	2004	2005
Pro-Kopf-Verbrauch (in kg)	9.2	9.5	9.7	9.6	9.9	10.5	10.8

1. Ermitteln Sie die entsprechende lineare Trendfunktion.
2. Prognostizieren Sie den Verbrauch für die Jahre 2006 und 2007.

Lösung:

Arbeitstabelle

Jahr	x_i	y_i	x_i^2	$x_i \cdot y_i$
1999	-3	9.2	9	-27.6
2000	-2	9.5	4	-19.0
2001	-1	9.7	1	-9.7
2002	0	9.6	0	0.0
2003	1	9.9	1	9.9
2004	2	10.5	4	21.0
2005	3	10.8	9	32.4
	0	69.2	28	7.0

1.

$$a_0 = \frac{692}{7} = 9.885, \qquad a_1 = \frac{7}{28} = 0.25\,,$$

d.h.

$$y^* = 9.885 + 0.25x\,.$$

2.

$$y^*_{2006} = y^*(4) = 10.9\,, \qquad y^*_{2007} = y^*(5) = 11.1\,.$$

D. 6. 4. (*Quasilineare Funktion*)
Eine Funktion heißt *quasilinear*, wenn sie sich in folgender Form darstellen lässt:

$$y^* = a_0 + a_1 \cdot F_1(x) + a_2 \cdot F_2(x) + ... + a_s \cdot F_s(x)\,.$$

BS. 6. 5.
1. Folgende Funktionen sind quasilinear:

(1) $y^* = a_0 + a_1 \cdot \frac{1}{x}$,

(2) $y^* = a_0 + a_1 x + a_2 x^2$.

2. Folgende Funktionen sind nicht quasilinear:

(1) $y^* = a_0 x^{a_1}$,

(2) $y^* = a_0 \cdot a_1{}^x$,

(3) $y^* = \frac{1}{a_0 + a_1 x}$

B. 6. 6.
Die *Normalgleichungen für eine quasilineare Funktion* lauten:

$$\begin{aligned}
n \cdot a_0 \quad &+ a_1 \cdot \sum_i F_1(x_i) \quad + a_2 \cdot \sum_i F_2(x_i) + \quad \ldots \quad = \sum_i y_i \\
a_0 \cdot \sum_i F_1(x_i) &+ a_1 \cdot \sum_i F_1^2(x_i) \quad + a_2 \cdot \sum_i F_1(x_i) \cdot F_2(x_i) + \ \ldots = \sum_i y_i \cdot F_1(x_i) \\
a_0 \cdot \sum_i F_2(x_i) &+ a_1 \cdot \sum_i F_1(x_i) \cdot F_2(x_i) + a_2 \cdot \sum_i F_2^2(x_i) + \quad \ldots \quad = \sum_i y_i \cdot F_2(x_i)
\end{aligned}$$

.
.
.

BS. 6. 6.
Es soll sie Abhängigkeit der Stückkosten von der hergestellten Stückzahl untersucht werden. In 15 Betrieben wurden die Daten ermittelt:

Betrieb	Hergestellte Stückzahl (1000)	Stückkosten (€)
1	2	8
2	3	10
3	4	7
4	4	6
5	5	5
6	6	5
7	6	4
8	6	3
9	7	4
10	8	5
11	9	3
12	10	2
13	12	1
14	13	1
15	14	2

Bestimmen Sie die entsprechenden Regressionsfunktionen in der Form

(1) $y^* = a_0 + a_1 \cdot \frac{1}{x}$,

(2) $y^* = a_0 + a_1 x + a_2 x^2$.

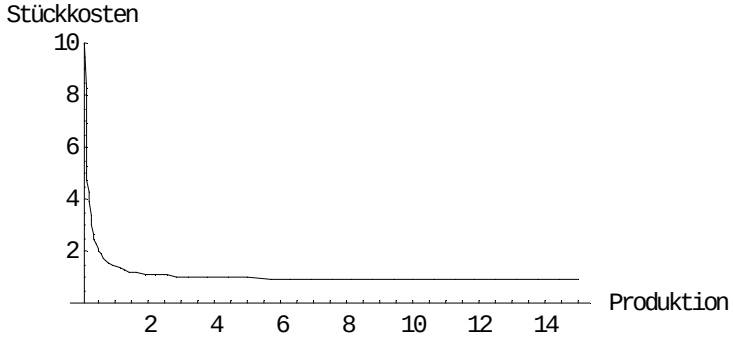
Lösung:
(1)

$$\begin{cases} n \cdot a_0 \quad + a_1 \sum_i \frac{1}{x_i} \quad = \sum_i y_i \\ a_0 \cdot \sum_i \frac{1}{x_i} + a_1 \cdot \sum_i \frac{1}{x_i^2} = \sum_i \frac{y_i}{x_i} \end{cases}$$

$$\sum_i \frac{1}{x_i} = 2.7455, \quad \sum_i \frac{1}{x_i^2} = 0.6862, \quad \sum_i y_i = 66, \quad \sum_i \frac{y_i}{x_i} = 15.6226 .$$

$$\begin{cases} 15a_0 + 2.7455a_1 = 66 \\ 2.7455a_0 + 0.6862a_1 = 15.6226 \end{cases} \quad \Rightarrow \quad a_0 = 0.8701, \quad a_1 = 19.286 ,$$

$$y^* = 0.8701 + 19.2855 \cdot \frac{1}{x}$$



(2)

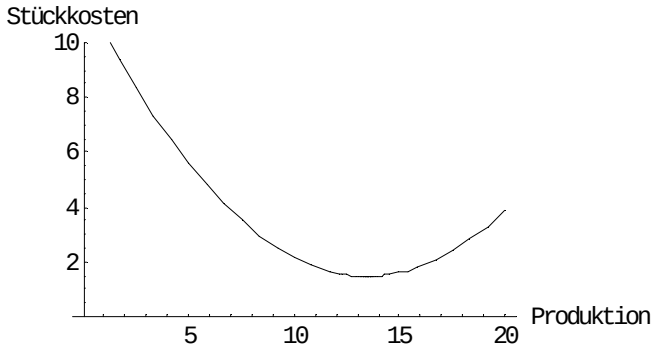
$$\begin{cases} n \cdot a_0 + a_1 \cdot \sum_i x_i + a_2 \cdot \sum_i x_i^2 = \sum_i y_i \\ a_0 \cdot \sum_i x_i + a_1 \cdot \sum_i x_i^2 + a_2 \cdot \sum_i x_i^3 = \sum_i x_i \cdot y_i \\ a_0 \cdot \sum_i x_i^2 + a_1 \cdot \sum_i x_i^3 + a_2 \cdot \sum_i x_i^4 = \sum_i x_i^2 \cdot y_i \end{cases}$$

$$\sum_i x_i = 109, \quad \sum_i x_i^2 = 981, \quad \sum_i x_i^3 = 10189, \quad \sum_i x_i^4 = 115893 ,$$

$$\sum_i y_i = 66, \quad \sum_i x_i \cdot y_i = 363, \quad \sum_i x_i^2 \cdot y_i = 2551 .$$

$$\Rightarrow \quad a_0 = 11.87, \quad a_1 = -1.541, \quad a_2 = 0.0571 .$$

$$y^* = 11.87 - 1.541x + 0.0571x^2 .$$



D. 6. 6. (*Bestimmtheitsmaß*)
Als *Bestimmtheitsmaß* im Allgemeinen bezeichnet man:

$$r^2_{yx} := \frac{\sum_{i=1}^{n}\left(y_i^* - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \qquad 0 \le r^2_{yx} \le 1.$$

.

BS. 6. 6. (*Fortsetzung*)

			Hyperbel		Parabel		
i	x_i	y_i	y_i^*	$(y_i^* - \bar{y})^2$	y_i^*	$(y_i^* - \bar{y})^2$	$(y_i - \bar{y})^2$
1	2	8	10.5131	37.3700	9.0164	21.3111	12.96
2	3	10	7.2988	8.4028	7.7609	11.2956	31.36
3	4	7	5.6916	1.6682	6.6196	4.9266	6.76
4	4	6	5.6916	1.6682	6.6196	4.9266	2.56
5	5	5	4.7273	0.1071	5.5925	1.4221	0.36
6	6	5	4.0844	0.0996	4.6796	0.0782	0.36
7	6	4	4.0844	0.0996	4.6796	0.0782	0.16
8	6	3	4.0844	0.0996	4.6796	0.0782	1.96
9	7	4	3.6252	0.6002	3.8809	0.2695	0.16
10	8	5	3.2809	1.2525	3.1964	1.4487	0.36
11	9	3	3.0130	1.9238	2.6261	3.1467	1.96
12	10	2	2.7987	2.5642	2.17	4.9729	5.76
13	12	1	2.4773	3.6969	1.6004	7.8378	11.56
14	13	1	2.3536	4.1876	1.4869	8.4862	11.56
15	14	2	2.2477	4.6325	1.4876	8.4821	5.76
	109	66		68.3729		78.7604	93.6

$\bar{y} = 4.4$

(1)

$$r^2 = \frac{68.3729}{93.6} = 0.7304797008547 \approx 0.73$$

d.h. die Stückkosten werden zu etwa 73% durch die Stückzahl bestimmt.

(2)

$$r^2 = \frac{78.7604}{93.6} = 0.84145726495727 \approx 0.84$$

d.h. die Stückkosten werden zu etwa 85% durch die Stückzahl bestimmt.

BS. 6. 7. (*Zur Linearisierung einer nichtlinearen Regressionsfunktion*)
In einem Land hat sich der Bestand an Rindern wie folgt entwickelt:

Jahr	1993	1995	1996	1999	2001	2005
Rinderzahl (in 10^5 Stück)	11.0	13.3	14.6	19.6	23.7	33.1

1. Stellen Sie die Entwicklung der Rinderzahl in Form der Regressionsfunktion

$$y^* = a_0 \cdot a_1{}^x .$$

2. Prognostizieren Sie den Bestand an Rindern im Jahre 2006.

Lösung:
1.

$$y^* = a_0 \cdot a_1{}^x ,$$
$$\log y^* = \log a_0 \cdot a_1{}^x , \quad \log y^* = \log a_0 + \log a_1{}^x , \qquad \log y^* = \log a_0 + x_1 \cdot \log a_1 .$$

Mit

$$Y^* := \log y^*, \quad A_0 := \log a_0, \quad A_1 := \log a_1$$

ergibt sich:

$$Y^* = A_0 + A_1 \cdot x$$

mit den Normalgleichungen

$$n \cdot A_0 \quad + A_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$
$$A_0 \cdot \sum_{i=1}^{n} x_i + A_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot Y_i$$

Arbeitstabelle

Jahr	x_i	y_i	$\lg y_i$	x_i^2	$x_i \cdot \lg y_i$
1993	1	11.0	1.04139	1	1.04139
1995	3	13.3	1.12385	9	3.37155
1996	4	14.6	1.16435	16	4.65740
1999	7	19.6	1.29226	49	9.04582
2001	9	23.7	1.37475	81	12.37275
2005	13	33.1	1.51983	169	19.75779
Summen	37	115.3	7.51643	325	50.24670

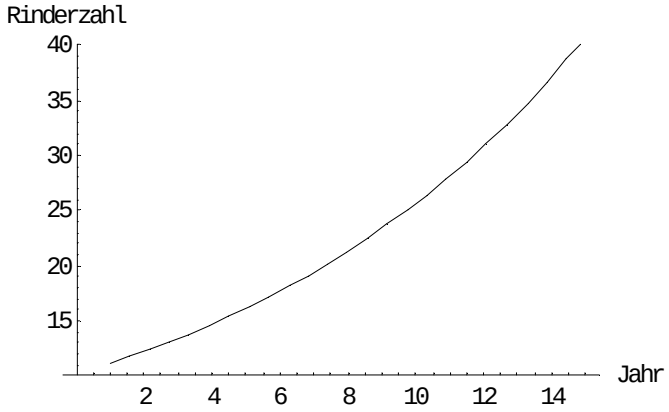
$$\begin{aligned} 6A_0 + 37A_1 &= 7.52 \\ 37A_0 + 325A_1 &= 50.25 \end{aligned} \quad \Rightarrow \quad A_0 = 1.004654392, \quad A_1 = 0.040034423 .$$

Damit erhält man

$$a_0 = 10.14972772, \quad a_1 = 1.096565109 .$$

Die gesuchte Regressionsfunktion lautet:

$$y^* = 10.149 \cdot 1.097^x .$$



2.

$$y^*(14) = 10.149 \cdot 1.097^{14} \approx 37.1 \ (10^5 \text{ Stück}) .$$

BS. 6. 8.

Stellen Sie die Normalgleichungen zur Bestimmung der Regressionsfunktion

$$y^* = a_0 \cdot a_1{}^x$$

auf.

Lösung:

$$S(a_0, a_1) = \sum_i \left(y_i - a_0 \cdot a_1{}^{x_i} \right)^2 \rightarrow Min!$$

$$S_{a_0}(a_0, a_1) = -2 \cdot \sum_i \left(y_i - a_0 \cdot a_1^{x_i} \right) \cdot a_1^{x_i}$$
$$S_{a_1}(a_0, a_1) = -2 \cdot \sum_i \left(y_i - a_0 \cdot a_1^{x_i} \right) \cdot a_0 \cdot x_i \cdot a_1^{x_i - 1}$$

$$\begin{cases} -2 \cdot \sum_i \left(y_i - a_0 \cdot a_1^{x_i} \right) \cdot a_1^{x_i} = 0 \\ -2 \cdot \sum_i \left(y_i - a_0 \cdot a_1^{x_i} \right) \cdot a_0 \cdot x_i \cdot a_1^{x_i - 1} = 0 \end{cases}$$