Regression und Korrelation (Gurtner)

Oft muss man Trends verlängern. Dazu nimmt man die Daten einer Entwicklung, legt eine passende Kurve durch und verlängert die Kurve in die Zukunft. Hier verwenden wir nur die Gerade als Kurve – das nennt sich dann **Regressionsgerade** oder Ausgleichsgerade (in EXCEL: Trendkurve)

Beispiel:
Folgender (fiktiver) Zusammenhang zwischen den Geburtszahlen und der Anzahl der Störche in Rust möge für verschiedene Jahre gegeben sein:

Störche	8	7	5	3	4
Geburten	12	10	8	5	6

Untersuchen Sie, ob ein linearer Zusammenhang zwischen der Anzahl der Geburten und der Anzahl der Störche besteht, um die These zu belegen, dass die Störche die Babys bringen (!)
Berechnen Sie die Korrelation und die Regressionsgrade und bestimmen Sie damit näherungsweise für 10 Störche die Anzahl der Geburten.
Zeichnen Sie für die Daten und die Gerade ein Diagramm!

Lösung:
Dazu müssen wir eine Tabelle erstellen, welche die Daten und deren Quadrate und Produkte enthält und zum Schluss die Summen und Mittelwerte:

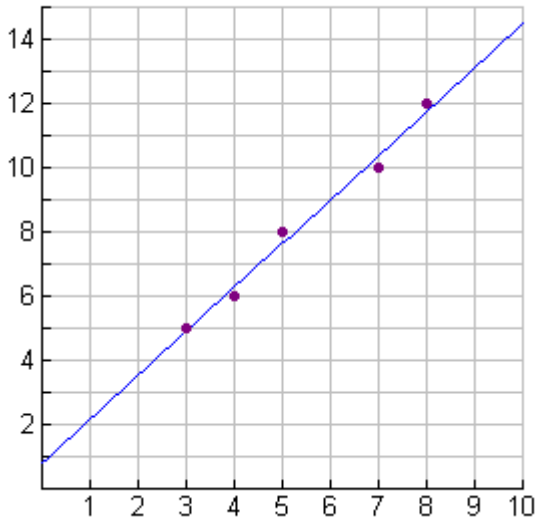
	xi	**yi**	**xi²**	**xi*yi**	**yi²**
	8	12	64	96	144
	7	10	49	70	100
	5	8	25	40	64
	3	5	9	15	25
	4	6	16	24	36
Summen	27	41	163	245	369
Mittelwert (:5)	5,4	8,2	32,6	49	73,8

Dadurch lassen sich die **Varianzen** von x und y und die **Kovarianz** berechnen:
V(x) = σ_x^2 = Mittelwert der x-Quadrate minus Quadrat des x-Mittelwerts = 32,6 – 5,4² = **3,44**
V(y) = σ_y^2 = Mittelwert der y-Quadrate minus Quadrat des y-Mittelwerts = 73,8 – 8,2² = **6,56**
coV(xy) = σ_{xy} = Mittelwert der Produkte minus Produkt der Mittelwerte = 49 – 5,4*8,2 = **4,72**

Damit kann man die Werte k und d der Regressionsgeraden bestimmen:
k = coV(xy)/ V(x) = 4,72 / 3,44 **= 1,37**
d = $\bar{y} - k \cdot \bar{x}$ = 8,2–1,372*5,4 = **0,79**
➔ **y = 1,37•x + 0,79** ist die Regressionsgerade.
Um die **Gerade einzeichnen** zu können, bestimmen wir die Werte am Beginn und Ende der Tabelle:
für x = 3 gilt: y = 1,37•3+ 0,79 = 4,9
für x = 8 gilt: y = 1,37•8 + 0,79 = 11,7
Für die **Korrelation** gilt:

r = coV(xy) / Wurzel aus V(x)*V(y) = $\frac{4,72}{\sqrt{3,44*6,56}}$ = 0,9936

und das Quadrat der Korrelation gibt **die Güte der Regressionsgeraden** an: r² = 0,897 = 89,7%.
Das ist gut, die Güte ist erst unter 50% so schlecht, dass man keine sichere Aussage mehr machen kann, ob die Daten korreliert sind oder nicht. Die Gerade liefert gute Prognosen!



Die Prognose für 10 Störche: y(10) = 1,37•10 + 0,79 = **14,5 Geburten**

Im Tascherechner TI–30 II ist das einfach + ohne Formeln einzutippen ➔ siehe Anhang!

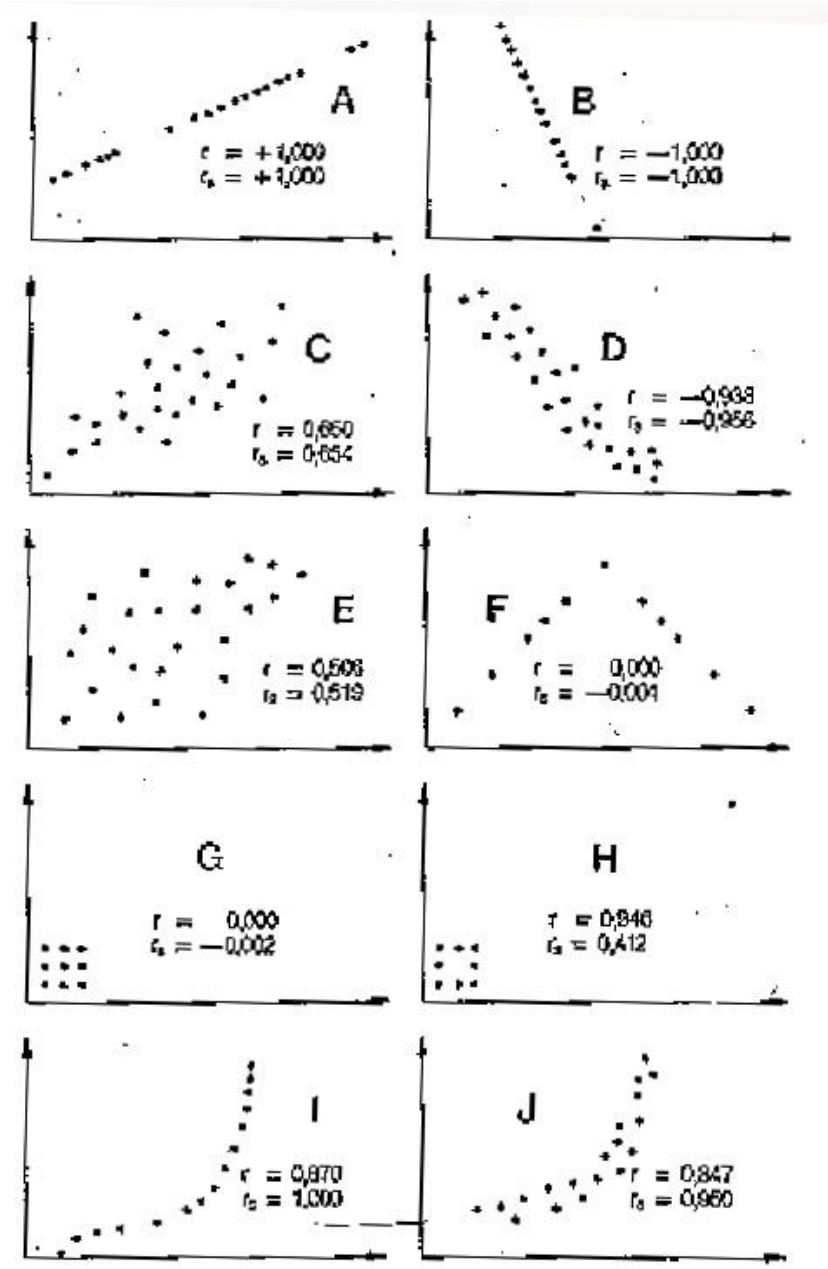
Was bedeuten nun Korrelation und Bestimmtheitsmaß der Güte?

Die Korrelation gibt an, wie gut die Daten zusammenhängen.
Positive Korrelationen heiße: Je mehr x desto mehr y ➔ Zeichnung A,C
Negative Korrelationen heißen: Je mehr x desto weniger y ➔ Zeichnung B,D

Korrelationen im Bereich zwischen –0,4 und 0,4 heißen, dass die Daten nur wenig korrelieren.
➔ Zeichnung E,F,G

Es könnten auch ganz andere Trägerfunktionen sinnvoll sein, wie Beispiel I und J zeigen. Da wäre eine quadratische Funktion weitaus besser angepasst.

Das **Bestimmtheitsmaß** gibt an, wie viel Prozent der Daten durch die Regressionsgerade gut dargestellt werden. Bei r = 0,5 ergibt sich $r^2 = 0,25$, also werden nur 25 % der Daten durch die Regressionsgerade richtig dargestellt.



Die ***Kovarianz***:

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}$$

("Mittelwert der Produkte minus Produkt der Mittelwerte")
gibt die Größe des linearen Zusammenhangs zweier Größe x und y an.
Bei der Methode der *linearen Regression* nimmt man an, dass zwischen den beiden Werten ein linearer Zusammenhang besteht, das heißt:

y = k•x + d + ein zufälliger Fehler

Die Konstanten k und d werden so bestimmt, dass die Summe der Quadrate der Fehler möglichst klein wird (Methode der kleinsten Fehlerquadrate von *C.F. Gauß*). Anschaulich können wir uns das so vorstellen, dass wir x und y als Koordinaten von Punkten auffassen und in ein Koordinatensystem einzeichnen. Wir suchen dann die Gerade, die diese Punktwolke am besten annähert. Diese Aufgabe kann man mit Hilfe der Differentialrechnung lösen und erhält als Gleichung der Regressionsgeraden: y = kx + d, wobei

$$a = \frac{\mathrm{Cov}(x,y)}{V(x)}$$

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

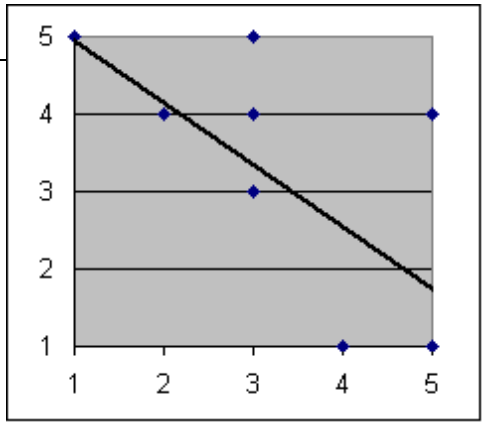
Die zweite Formel ergibt sich daraus, dass die Regressionsgerade durch den "Schwerpunkt" $(\bar{x}/\bar{y})$ der Punktwolke geht.
Der *Korrelationskoeffizient* r liefert ein Maß dafür, wie gut die gegebenen Werte durch diese lineare Funktion angenähert werden. Er ist definiert durch

$$r = \frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)\cdot V(y)}}$$

Beispiel 2
soll zeigen, wie es mit negativen Korrelationen aussieht: Der Zusammenhang von (fiktiven) Mathe- und Deutschnoten soll gefunden werden

	x=Mathenote	y=Englischnote	x*y	x²	y²	Ausgleichs–Näherung	Differenz zur Englischnote
	1	5	5	1	25	4,9	-0,1
	4	1	4	16	1	2,5	1,5
	3	4	12	9	16	3,3	-0,7
	3	5	15	9	25	3,3	-1,7
	2	4	8	4	16	4,1	0,1
	3	3	9	9	9	3,3	0,3
	5	1	5	25	1	1,7	0,7
	5	4	20	25	16	1,7	-2,3
	4	1	4	16	1	2,5	1,5
	3	3	9	9	9	3,3	0,3
Summe	**33**	**31**	**91**	**123**	**119**		
Mittelwerte (:10)	**3,3**	**3,1**	**9,1**	**12,3**	**11,9**		

Varianz von x	12,3 – 3,3² =	**1,41**
Varianz von y	11,9 – 3,1² =	**2,29**
Kovarianz von xy	9,1 – 3,3*3,1 =	**–1,13**
k = Cov(xy) / V(x)	–1,13 : 1,41 =	**– 0,80**
d = $\bar{y} - k\cdot\bar{x}$	3,1–(**-0,8**)*3,3 =	**5,74**
	r = –1,13 : √(1,41*2,29)	–0,63
	r² = (–0,63)²	0,40
	Ausgleichsgerade:	**y= –0,80*x + 5,74**



Wie man sieht ist die Korrelation negativ, je besser die Note in Mathematik ist , desto schlechter ist sie in Englisch. Diese Aussage trifft auf 40% der Daten zu. Das ist nicht sehr aussagekräftig, wie man an der Zeichnung sieht – die Punkte sind recht weit weg von der Geraden!

Übungen:

1) In einem Unternehmen werden für verschiedene Produktionsmengen die Kosten festgestellt:

Stückmenge	**Kosten**
30	128
35	165
40	175
55	240
65	300

Ermitteln Sie die Regressionsgerade und die Korrelation.
Wie groß ist die Differenz zwischen den realen Kosten bei 35 Stück und den berechneten?
Bei wie viel Stück sind die Kosten auf 400 gestiegen?

2) Die durchschnittlichen Rohölpreise am Spotmarkt Amsterdam in den letzten 8 Monaten waren (in $/Barrel)

Monat	**Preis**
3	28,4
4	29,0
5	30,3
6	30,6
7	29,8

Bestimmen Sie die Regressionsgerade und prognostizieren Sie den Ölpreis für die nächsten 2 Monate. Wie viel Prozent der Daten erklärt die Regressionsgerade?
In welchem Monat übersteigt der Preis 33 $?

3) Die folgenden Daten sollen den Zusammenhang zwischen dem Gewicht (Masse) und der Körpergröße belegen:

Körpergröße	**Masse**
164	48
169	68
160	51
171	54
165	53
165	66

Bestimmen Sie die Regressionsgerade und die Korrelation. Wie gut ist die Aussage: „Je größer desto mehr Gewicht“ dokumentiert?
Welchen Wert gibt die Regressionsgerade für x =165 cm? Welche Körpergröße sollte ein 100 kg-Mensch haben?

4) Das **Alter** von Ehepaaren ist gegeben:

Ehemann	Ehefrau
24	32
18	20
37	35
26	30
19	22
21	26
20	24
31	38

Kann man auf Grund der Regressionsgeraden von einem Zusammenhang: „Je älter der Mann, desto älter die Frau“ reden?
Zeichnen Sie die Punkte und die Gerade in eine Zeichnung.
Wie groß sind die Mittelwerte und die Standardabweichungen des Alters von Mann und Frau?

5) Folgende Tabelle zeigt die **Weizenproduktion** im Jahre 1982

Land	Fläche in Mio. ha	Produktion in Mio. t
Sowjetunion	57	87
USA	32	76
China	28	63
Indien	22	38
Kanada	13	28
Australien	12	9
Türkei	9	18
Frankreich	5	25
England	2	10
Deutschland	2	9

Zeichnen Sie ein Streudiagramm und die Regressionsgerade in eine Zeichnung. Welche Länder weichen am weitesten von der Geraden ab? Warum?
Welchen Wert für die Produktionsmenge würde man f ür ein Land mit 10 Mio. ha Anbaufläche erwarten?

6) Folgende Messwerte von Meereshöhe und Jahresmitteltemperatur sind gegeben:

Station	**Seehöhe**	**Temp.**
Berlin	49	9,1
Prag	374	7,9
Wien	203	9,1
Feuerkogel	1592	3,3
Salzburg	437	8,6
Budapest	130	10,9
Zugspitze	2962	-5,0
Innsbruck	579	8,4
Säntis	2496	-2,3
Sonnblick	3106	-6,4

Ermitteln Sie die Regressions-gerade und das Streudiagramm.
Wie groß wird demnach die Temperatur in 800 m Meereshöhe sein? Welche Meereshöhe sollte man bevorzugen, wenn man ein Jahresmittel von 15° haben will?

7) Eine Datenliste des **PKW-Bestandes** von Österreich ist gegeben:

Jahr	1950	1955	1960	1965	1970
PKW in 1000 Stk.	**51**	**143**	**404**	**791**	**1197**

Machen Sie eine Prognose für 1975-1990 und vergleichen Sie mit den realen Daten:

Jahr	1975	1980	1985	1990
PKW in 1000 Stk.	**1721**	**2247**	**2531**	**2991**

8) Durch Messung des Wasserablaufs im **Kaffeefilter** einer Kaffeemaschine mit einer Filtertiefe von 10cm ergibt sich folgende Tabelle:
Wann wird der Wasserstand 0 sein und der Kaffee fertig sein?

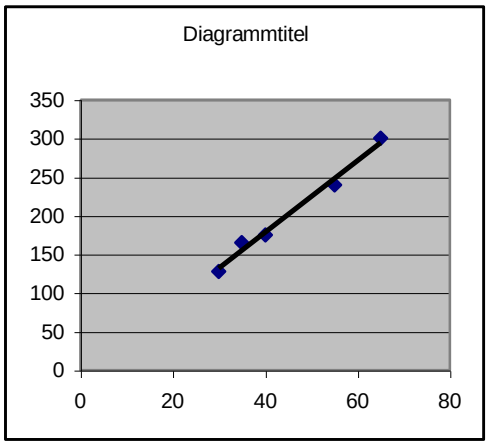
Zeit in sec	Wasserstand in cm
0	10
10	8
20	6
33	4
50	2

9) Der **Stromverbrauch** in Österreich in den Jahren 1975-1980 ist in der folgenden Tabelle gegeben. Legen Sie eine Regressionsgerade durch diese Daten und ermitteln Sie die Verbrauchswerte für 1981-1983

Jahr	Stromverbrauch in Md. kWh
1975	30,7
1976	33,1
1977	33,7
1978	35,3
1979	36,8
1980	38

Lösungen:

	Stückmenge	Kosten	x*y	x²	y²
	30	128	3840	900	16384
	35	165	5775	1225	27225
	40	175	7000	1600	30625
	55	240	13200	3025	57600
	65	300	19500	4225	90000
Summe	225	1008	49315	10975	221834
durch 5	**45**	**201,6**	9863	2195	44367
	= xq	**= yq**			
quadriert und verschoben			9072	2025	40643
Differenz			**791**	**170**	**3724,2**
Wurzel			**= sxy**	**13,04**	**61,03**
				=sx	**= sy**



Geradensteigung k= sxy : sx² =	4,65
Geradenabstand d= yq - k*xq =	-7,78
Korrelation r = sxy : (sx*sy) =	0,99
Bestimmtheitsmaß r² =	0,99

1) Die Regressionsgerade ist y= 4,65*x – 7,78, für x= 35 ergibt das y= 4,65*35-7,78 = 155
Die Differenz zu den realen Kosten ist: 10. Bei 88 Stück sind die Kosten ca. 400 GE

2) y = 0,44 x + 27,42 r = 0,76 r² = 0,58 (58%=„Genügend“) Ölpreise: (8→30,94$) (9→31,38$)
Im 12. Monat übersteigt der Preis die 32$-Marke

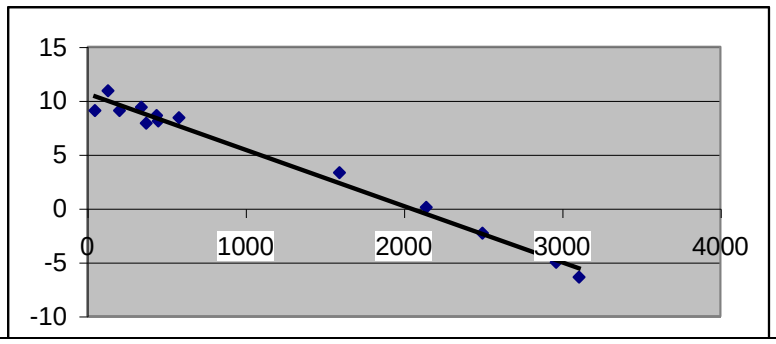
3) y = 0,88 x – 89,2 r = 0,41 r² = 0,17 (wenig!) x=165→y= 56 kg; y=100 kg→215 cm!!

4) y = 0,86 x + 7,28 r = 0,88 r² = 0,79 Zusammenhang ist eher verkehrt quadratisch!
x= 24,5 ±6,2 y= 28,3 ±5,9 (Mittelwert± Standardabweichung)

5) y = 1,57 x + 7,6 r = 0,93 r² = 0,87 Australien unten (Wüste?) USA oben (no na!) 23,3 Mio.t

6) y = -0,00525 x +10,63 r = -0,99 r² = 0,985 800m→6,4° 15°→-832 m (unter dem Meeresspiegel)

1592	3,3
437	8,6
130	10,9
2962	-5
579	8,4
2496	-2,3
3106	-6,4
342	9,4
448	8,1
2140	0,1



7) y = 58,8 x – 3010,8 r = 0,98 r² = 0,95

Jahr	1975	1980	1985	1990
PKW in 1000 Stk. real	**1721**	**2247**	**2531**	**2991**
Prognose mit Gerade	**1399**	**1693**	**1987**	**2281**

8) y = –0,16 x + 9,62 r = –0,99 r² = 0,98 Wasserstand 0 nach 60 sec

9) y =1,4 x + 31 Verbrauchswerte 1981: 39,5, für 1982: 40,9 für 1983: 42,3 (real war es weniger!)

Anhang:

Berechnung der Korrelation und Regression mit Taschenrechner TI–30 II:

Wir wollen folgende Daten eingeben:

x	2	3	4
y	1	2	6

Zum Start wollen wir eine etwaige vorige Berechnung löschen: 2nd EXIT STAT
Dann starten wir die Statistik–Eingabe mit 2nd STAT und geben **2–VAR** =
für die Statistik mit 2 Variablen ein.

Nun kommen die Daten. Dazu starten wir mit DATA
Der Reihe nach geben wir nun die Daten ein und wechseln die Zeile mit der Pfeil–nach–unten–Taste ⇩

$X_1 = 2$ ⇩ $Y_1 = 1$ ⇩ $X_2 = 3$ ⇩ $Y_2 = 2$ ⇩ $X_3 = 4$ ⇩ $Y_3 = 6$

Dann können wir schon die Ergebnisse sichten :

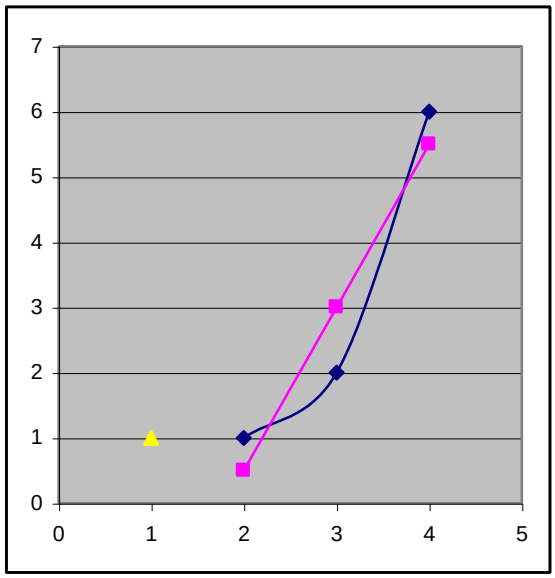
Mit STATVAR und der Pfeiltaste–nach–rechts ⇨ kommt man auf folgende Ergebnisse:

$n = 3$ $\bar{x} = 3$ $Sx = 1$ $\sigma x = 0{,}81$ $\bar{y} = 3$ $Sy = 2{,}64$ $\sigma y = 2{,}16$

und ganz hinten kommt unsere **Regresion** : $a = 2{,}5$ $b = -4{,}5$ $r = 0{,}94$
und mit der x^2–Taste bekommt man auch $r^2 = 0{,}89$ (89%–Güte)

Das ergibt die Ausgleichsgerade = Regressionlinie = **Trendgerade y = 2,5x – 4,5**

x	2	3	4
y	1	2	6
y–Trend: 2,5x–4,5	0,5	3	5,5



Zum Schluss wollen wir noch alles löschen: 2nd EXIT STAT