Übung 5

-

Zeitreihen, Regression und Korrelation

Musterlösung

Aktuelle Version: 19. Juli 2022

Hinweise:

- Übungen sind mit Vorteil alleine zu lösen.
- Benutzen Sie die Musterlösungen nur zur Korrektur.
- Die Übungen sind wichtige Vorbereitungen für die Prüfung. Lösen sie die Übungen sorgfältig und stellen Sie die Lösungswege übersichtlich dar.
- (Ergänzte) Vorlesungsunterlagen und Fachbücher helfen beim Lösen von Übungen und bringen gleichzeitig eine erweiterte Ansicht auf die Problemstellung.
- Wenn Sie die Übungen nicht verstehen, fragen Sie!

Übung 1. *Fragen*

1. Was ist eine Zeitreihe?

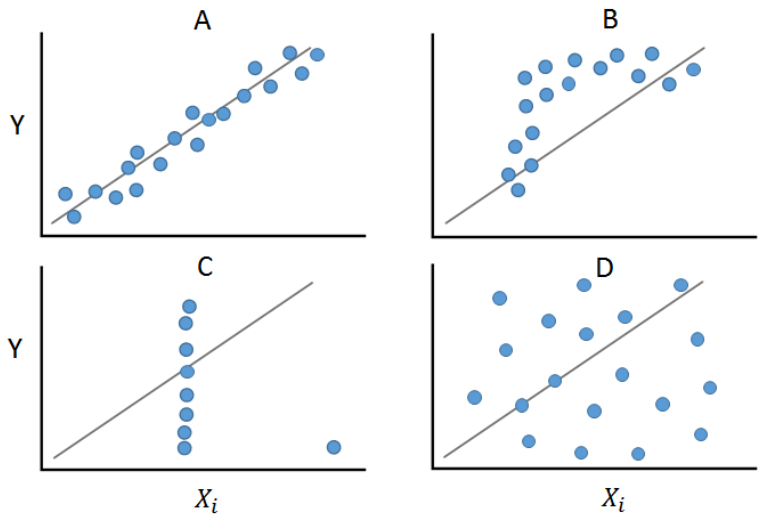
   Eine zeitlich geordnete Folge von Merkmalswerten.

2. Welche Fragen interessieren bei der Untersuchung des Zusammenhangs zwischen zwei Merkmalen?

   Ob zwischen Ihnen ein rechnerischer Zusammenhang besteht. Das heisst z.B., steigt Wert A, wenn Wert B steigt?

3. Wozu kann ein Streudiagramm in einem ersten Analyseschritt genutzt werden?

   Das Streudiagramm gibt einen ersten Überblick über den Zusammenhang zweier Merkmale. Deutet sich z.B. ein linearer (A), ein exponentieller (B) oder kein Zusammenhang (C, D) an?[1]



4. Wir haben die lineare Regression mit der Methode der kleinsten Quadrate kennengelernt. Welche weiteren, sinnvollen Formen können sie sich vorstellen?

   Möglich wäre z.B. eine graphische Lösung, indem von Auge die Regressionsgerade in das Streudiagramm eingepasst wird.

5. Erläutern Sie den Begriff Kovarianz!

   Die Kovarianz ist ein Mass für den Grad des miteinander Variierens der Messwertreihen zweier Variablen.

6. Erläutern Sie den Begriff Korrelationskoeffizient! Welchen Wertebereich kann der Korrelationskoeffizient annehmen?

   Der Korrelationskoeffizient beschreibt die Enge des linearen Zusammenhangs zweier Merkmale durch eine Zahl r, die zwischen $+1$ und -1 liegt. Bei $r = 1$ sprechen wir von einem perfektiven positiven (linearen) Zusammenhang, bei $r = -1$ von einem perfekten negativen (linearen) Zusammenhang, bei $r = 0$ von keinem linearen Zusammenhang. $r = -1$

[1] Bildquelle `https://wikis.fu-berlin.de/display/fustat/Residuenplots`

7. Können Sie davon ausgehen, dass bei einem Korrelationskoeffizient von $|r| < 1$ *kein* perfekter Zusammenhang besteht?

   Nein, es könnte einen perfekten nichtlinearen Zusammenhang geben. Ebenso kann es bei einem Korrelationskoeffizient von $r = 0$ einen nichtlinearen Zusammenhang geben.

8. Welcher Skala müssen Ihre Daten genügen, um eine Bravais-Pearson-Korrelation zu berechnen?

   Die Daten müssen mindestens intervallskaliert sein.

9. Welche Kausalität liegt bei einem Korrelationskoeffizient $|r| > 0$ der Merkmale x und y vor?

   Die Korrelation gibt keine Auskunft über Kausalität, möglich wäre:

   (a) x beeinflusst y
   (b) y beeinflusst x
   (c) x und y beeinflussen sich gegenseitig
   (d) ein drittes Merkmal z beinflusst sowohl x als auch y

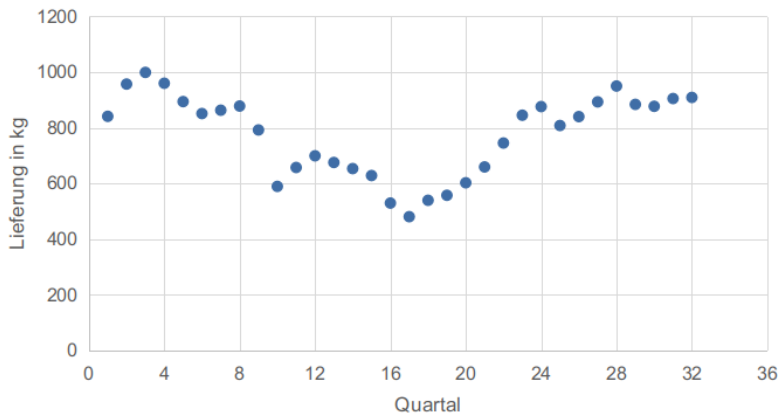
10. Erklären Sie den Unterschied zwischen formaler und sachlicher Abhängigkeit!

    Eine Formale Abhängigkeit zwischen zwei Datensätzen ist gegeben, wenn die Korrelation ungleich 0 ist. Eine kausale oder sachliche Abhängigkeit lässt sich daraus nicht automatisch ableiten. Die formale Abhängigkeit kann aber ein Indiz für einen Zusammenhang sein und ist durch einen Domänen-Experten zu bewerten.

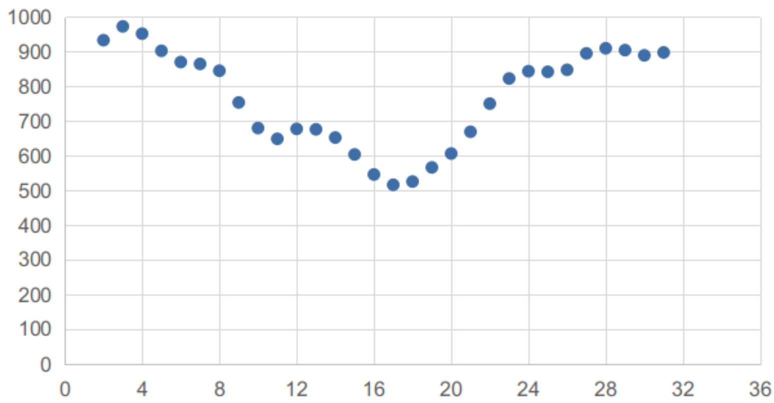
Übung 2. *Trend*
Die Daten der vierteljährlichen Anlieferungen in kg von Schlacke sind unten angeben:

$$y_i = (841, 957, 999, 960, 894, 851, 863, 878, 792, 589, 657, 699, 675, 653, 628, 529, 480, 539, 557, 602, 659, 745, 845, 876, 808, 840, 893, 950, 884, 877, 905, 909)$$

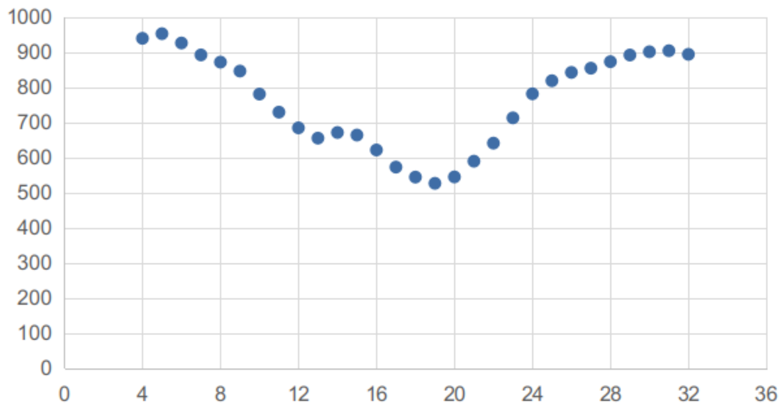
1. Stellen Sie den Verlauf der Zeitreihe zeitlich dar.



2. Stellen Sie den Verlauf des gleitenden Mittels über 3 Quartale dar.



3. Stellen Sie den Verlauf des gleitenden Mittels über 4 Quartale dar.



4. Können sie einen Trend erkennen?

   Bis zum 5. Jahr der Messungen wurde die gelieferte Schlacke immer weniger. Ab dem 5. Jahr steigt das Gewicht wieder, kommt aber im 7. und 8. Jahr langsam zum Stagnieren.

Übung 3. *Regression*
Der erforderliche Verbrauch an Grundmaterial eines Produktionsprozesses ist abhängig von der eingestellten Grösse des Produktes. Die Grösse kann jeden beliebigen reellen Wert zwischen 1 und 7 Einheiten einnehmen.

Um eine Formel zu entwickeln, die den Bedarf an Grundmaterial als Funktion der Grösse abschätzt, sind Experimente durchgeführt worden.

Die im Experiment gemessenen Daten sind jeweils für fünf Produktionstypen unten tabellarisch erfasst (Typ 1, Typ 2, Typ 3, Typ 4, Typ 5).

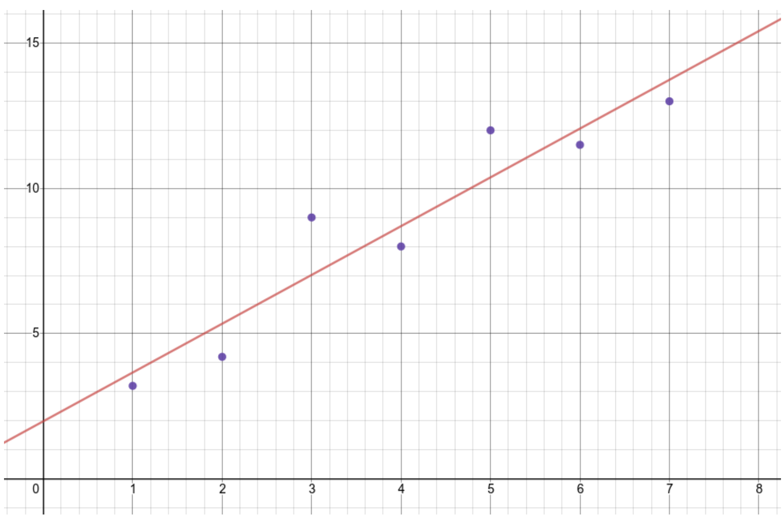
Bestimme Sie für jeden der Produktionstypen:

- Das entsprechende Streudiagramm und dadurch den vermuteten Zusammenhang.
- Eine Formel, die den Wert f(x) bestimmt. Visualisieren Sie die Formel im Streudiagramm.
- Die Kovarianz und den Korrelationskoeffizienten, sofern sinnvoll.

1. Typ 1

x_i	y_i
1	3.2
2	4.2
3	9
4	8
5	12
6	11.5
7	13

Das Streudiagramm deutet auf einen linearen Zusammenhang hin:



Wir berechnen die lineare Regression:

x_i	y_i	$x_i y_i$	x_i^2
1	3.2	3.2	1
2	4.2	8.4	4
3	9	27	9
4	8	32	16
5	12	60	25
6	11.5	69	36
7	13	91	49
28	60.9	290.6	140

$$n = 7$$
$$\bar{x} = \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4$$
$$\bar{y} = \frac{\sum_1^n y_i}{n} = \frac{60.9}{7} = 8.7$$
$$b = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{290.6 - 7 \cdot 4 \cdot 8.7}{140 - 7 \cdot 4^2} = 1.68$$
$$a = \bar{y} - b\bar{x} = 8.7 - 1.68 \cdot 4 = 1.98$$

Und erhalten die Funktion $f(x) = a + bx = 1.98 + 1.68x$. Die Funktion passt zu den Datenpunkten, siehe Diagramm oben.

Wir berechnen den Korrelationskoeffizient mit Hilfe der individuellen Varianzen und der Kovarianz:

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3.2	-3	9	-5.5	30.25	16.5
2	4.2	-2	4	-4.5	20.25	9
3	9	-1	1	0.3	0.09	-0.3
4	8	0	0	-0.7	0.49	0
5	12	1	1	3.3	10.89	3.3
6	11.5	2	4	2.8	7.84	5.6
7	13	3	9	4.3	18.49	12.9
28	60.9		28		88.3	47

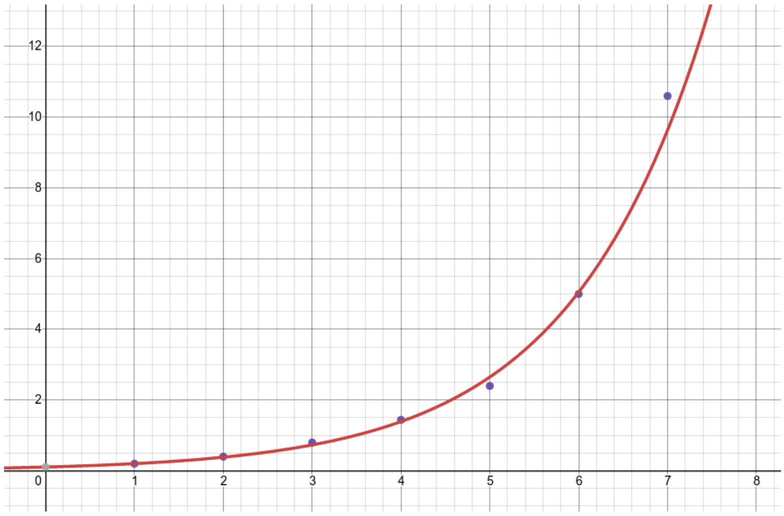
$$\sigma_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{47}{7} = 6.71$$
$$\sigma_x = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{28}{7}} = 2$$
$$\sigma_y = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{88.3}{7}} = 3.55$$
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{6.71}{2 \cdot 3.55} = 0.95$$

Der Korrelationskoeffizient von $r = 0.95$ verdeutlicht nochmals den starken positiven Zusammenhang.

2. Typ 2

x_i	y_i
1	0.20
2	0.40
3	0.80
4	1.44
5	2.40
6	5.00
7	10.60

Das Streudiagramm deutet auf einen exponentiellen Zusammenhang hin:



Wir logarithmieren die y-Werte $y' = ln(y)$ und berechnen anschliessend die lineare Regression:

x_i	y_i	y'_i	$x_i y'_i$	x_i^2
1	0.2	-1.61	-1.61	1
2	0.4	-0.92	-1.84	4
3	0.8	-0.22	-0.66	9
4	1.44	0.36	1.44	16
5	2.4	0.88	4.40	25
6	5	1.61	9.66	36
7	10.6	2.36	16.52	49
28		2.46	27.91	140

$$n = 7$$
$$\bar{x} = \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4$$
$$\bar{y}' = \frac{\sum_1^n y'_i}{n} = \frac{2.46}{7} = 0.35$$
$$b = \frac{\sum_1^n x_i y'_i - n\bar{x}\bar{y}'}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{27.91 - 7 \cdot 4 \cdot 0.35}{140 - 7 \cdot 4^2} = 0.65$$
$$a = \bar{y}' - b\bar{x} = 0.35 - 0.65 \cdot 4 = -2.25$$

Und erhalten die Funktion $f(x) = e^{a+bx} = e^{-2.25+0.65x}$. Die Funktion passt zu den Datenpunkten, siehe Diagramm oben.

Wir berechnen den Korrelationskoeffizient der *logarithmierten Werte* mit Hilfe der individuellen Varianzen und der Kovarianz:

x_i	y_i	y'_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y'_i - \bar{y}'$	$(y_i - \bar{y}')^2$	$(x_i - \bar{x})(y_i - \bar{y}')$
1	0.2	-1.61	-3	9	-1.96	3.84	5.88
2	0.4	-0.92	-2	4	-1.27	1.61	2.54
3	0.8	-0.22	-1	1	-0.57	0.32	0.57
4	1.44	0.36	0	0	0.01	0	0
5	2.4	0.88	1	1	0.53	0.28	0.53
6	5	1.61	2	4	1.26	1.59	2.52
7	10.6	2.36	3	9	2.01	4.04	6.03
28		2.46		28		11.68	18.07

$$\sigma_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{18.07}{7} = 2.58$$

$$\sigma_x = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{28}{7}} = 2$$

$$\sigma_y = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{11.68}{7}} = 1.29$$

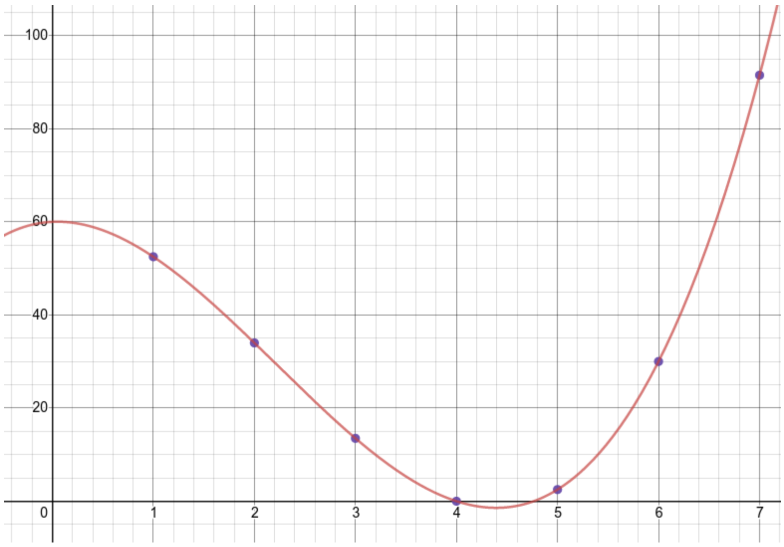
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{2.58}{2 \cdot 1.29} = 0.999$$

Der Korrelationskoeffizient von $r = 1.999$ verdeutlicht nochmals den beinahe perfekten positiven Zusammenhang.

3. Typ 3

x_i	y_i
1	52.5
2	34.0
3	13.5
4	0.0
5	2.5
6	30.0
7	91.5

Das Streudiagramm deutet auf einen polynomiellen Zusammenhang hin:



Wir verwenden den Newton-Algorithmus:

x_i	y_i	$D_{i,i-1}$	$D_{i,\ldots,i-2}$	$D_{i,\ldots,i-3}$	$D_{i,\ldots,i-4}$	$D_{i,\ldots,i-5}$	$D_{i,\ldots,i-6}$
1	52.5						
2	34	-18.5					
3	13.5	-20.5	-1				
4	0	-13.5	3.5	1.5			
5	2.5	2.5	8	1.5	0		
6	30	27.5	12.5	1.5	0	0	
7	91.5	61.5	17	1.5	0	0	0

$$a_0 = y_1 = 52.5$$
$$a_1 = D_{2,1} = -18.5$$
$$a_2 = D_{3,2,1} = -1$$
$$a_3 = D_{4,3,2,1} = 1.5$$
$$a_4 = D_{5,4,3,2,1} = 0$$
$$a_5 = D_{6,5,4,3,2,1} = 0$$
$$a_6 = D_{7,6,5,4,3,2,1} = 0$$

Und erhalten die das folgende Polynom. Die Funktion passt zu den Datenpunkten, siehe Diagramm oben.

$$f(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)(x - x_2) + ... + a_n(x - x_1)...(x - x_n)$$
$$f(x) = 52.5 - 18.5(x - 1) - 1(x - 1)(x - 2) + 1.5(x - 1)(x - 2)(x - 3)$$
$$f(x) = 1.5x^3 - 10x^2 + x + 60$$

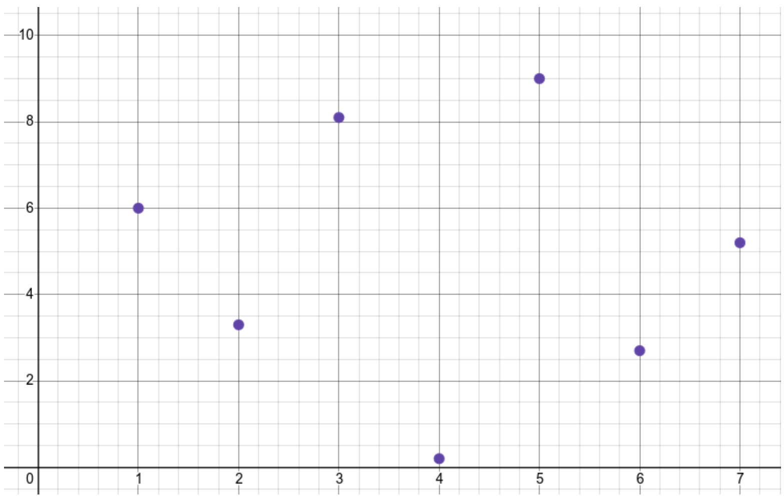
Bemerkenswert ist, dass alle Werte ab a_4 gleich null sind, somit ist das Polynom vom Grad 3. Da die Werte exakt 0 werden, ist die Zahlenfolge wohl direkt von einem Polynom abgeleitet worden (tatsächlich ist das auch so).

Wenn sie die Werte leicht variieren, kommen höhere Potenzen hinzu. Letztlich ist es auch eine Abschätzungsfrage, wie genau sie die Punkte verbinden wollen.

Korrelationskoeffizient ist nur für lineare Regression sinnvoll.

4. Typ 4

x_i	y_i
1	6
2	3.3
3	8.1
4	0.2
5	9
6	2.7
7	5.2



Aus dem Streudiagramm wird kein Zusammenhang deutlich. Eventuell könnte ein linearer Zusammenhang bestehen, wir berechnen daher den Korrelationskoeffizienten:

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-3	9	1.1	1.21	-3.3
2	3.3	-2	4	-1.6	2.56	3.2
3	8.1	-1	1	3.2	10.24	-3.2
4	0.2	0	0	-4.7	22.09	0
5	9	1	1	4.1	16.81	4.1
6	2.7	2	4	-2.2	4.84	-4.4
7	5	3	9	0.1	0.01	0.3
28	34.3		28		57.76	-3.3

Und erhalten:

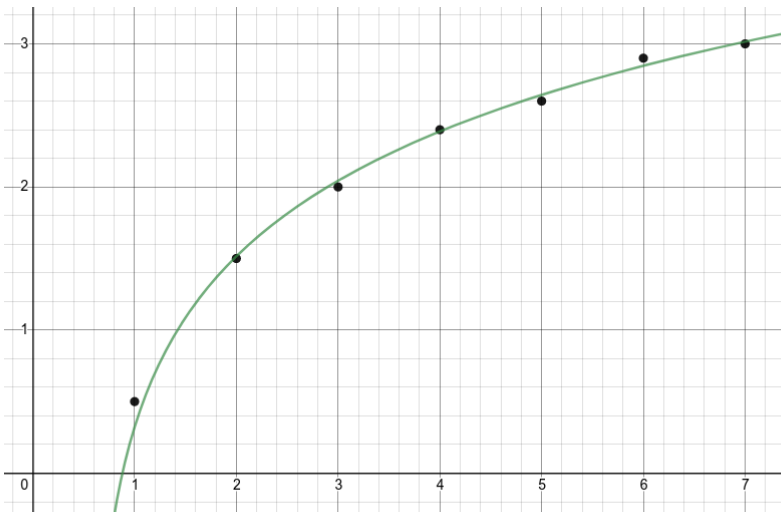
$$
\begin{aligned}
n &= 7\\
\bar{x} &= \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4\\
\bar{y} &= \frac{\sum_1^n y_i}{n} = \frac{34.3}{7} = 4.9\\
\sigma_{xy} &= \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{-3.3}{7} = .47\\
\sigma_x &= \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{28}{7}} = 2\\
\sigma_y &= \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{57.76}{7}} = 2.87\\
r &= \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{.47}{2 \cdot 2.87} = 0.08
\end{aligned}
$$

Der Korrelationskoeffizient von $r = 0.08$ deutet auf keinen (linearen) Zusammenhang hin.

5. Typ 5

x_i	y_i
1	0.5
2	1.5
3	2
4	2.4
5	2.6
6	2.9
7	3

Das Streudiagramm deutet auf einen logarithmischen Zusammenhang hin:



Wir können, wie bei der Exponentialfunktion, prinzipiell beliebige nichtlineare Transformationen auf die Eingangsgrösse anwenden und anschliessen die Regression durchführen! Wir exponieren daher die y-Werte $y' = e^y$ und berechnen anschliessend die lineare Regression:

x_i	y_i	y'_i	$x_i y'_i$	x_i^2
1	0.5	1.65	1.65	1
2	1.5	4.48	8.96	4
3	2	7.39	22.17	9
4	2.4	11.02	44.08	16
5	2.6	13.46	67.30	25
6	2.9	18.17	109.02	36
7	3	20.09	140.63	49
28		76.26	393.81	140

$$n = 7$$
$$\bar{x} = \frac{\sum_1^n x_i}{n} = \frac{28}{7} = 4$$
$$\bar{y}' = \frac{\sum_1^n y'_i}{n} = \frac{76.26}{7} = 10.89$$
$$b = \frac{\sum_1^n x_i y'_i - n\bar{x}\bar{y}'}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{393.81 - 7 \cdot 4 \cdot 10.89}{140 - 7 \cdot 4^2} = 3.17$$
$$a = \bar{y}' - b\bar{x} = 10.89 - 3.17 \cdot 4 = -1.79$$

Und erhalten die Funktion $f(x) = ln(a + bx) = ln(-1.79 + 3.17x)$. Die Funktion passt zu den Datenpunkten, siehe Diagramm oben.

Wir berechnen den Korrelationskoeffizient der *exponierten Werte* mit Hilfe der individuellen Varianzen und der Kovarianz:

x_i	y_i	y'_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y'_i - \bar{y}$	$(y'_i - \bar{y})^2$	$(x_i - \bar{x})(y'_i - \bar{y})$
1	0.5	1.65	-3	9	-9.24	85.38	27.72
2	1.5	4.48	-2	4	-6.41	41.9	12.82
3	2	7.39	-1	1	-3.50	12.25	3.5
4	2.4	11.02	0	0	0.13	0.02	0
5	2.6	13.46	1	1	2.57	6.60	2.57
6	2.9	18.17	2	4	7.28	53.00	14.56
7	3	20.09	3	9	9.2	84.64	27.6
28		76.26		28		283.79	88.77

$$\sigma_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{88.77}{7} = 12.68$$
$$\sigma_x = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{28}{7}} = 2$$
$$\sigma_y = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{283.79}{7}} = 6.37$$
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{12.68}{2 \cdot 6.37} = 0.995$$

Der Korrelationskoeffizient von $r = 0.995$ verdeutlicht nochmals den beinahe perfekten positiven Zusammenhang.

Zusatzaufgaben

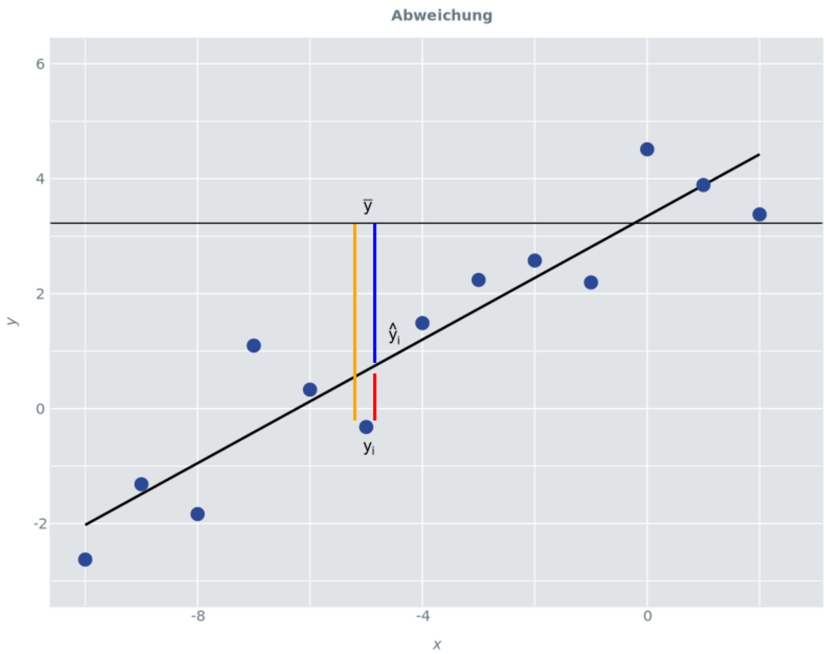
Übung 4. *Korrelations- und Determinationskoeffizienten*
In drei Untersuchungen mit gleich grossen Stichproben wurde jeweils eine einfache lineare Regression durchgeführt. Es wurden folgende Korrelationskoeffizienten ermittelt: $r_1 = 0.75$, $r_2 = 0.49$ und $r_3 = 0.62$.

1. Welche lineare Regression würden Sie verwenden?

   Hinweis: Der Determinationskoeffizient R^2 gibt an, wie viel der Gesamtvarianz durch die Regression erklärt wird[2]. Bei einer einfachen linearen Regression berechnet sich dieser mittels:

   $$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum e^2}{\sum(y_i - \bar{y})^2} = r^2$$

   Wobei es sich bei $\hat{y}$ um die durch die Regression vorhergesagten Werte, bei y um die tatsächlich gemessenen Werten handelt und bei $e = y - \hat{y}$ (im Bild rot) um das Residuum handelt[3]:



   Je besser eine Regression die Werte vorhersagt, desto kleiner sind die Absolutwerte der Residuen. D.h., je kleiner die Summe der Quadrate der Residuen, desto besser die Regression. Diese Summe wird minimal bei einem

[2] Mehr Informationen dazu findet sich zum Beispiel hier: https://www.inwt-statistics.de/blog-artikel-lesen/Bestimmtheitsmass_R2-Teil1.html

[3] Bildquelle https://www.inwt-statistics.de/blog-artikel-lesen/Bestimmtheitsmass_R2-Teil3.html

maximalem Determinationskoeffizient von $R^2 = r^2 = 1$, d.h. bei einem perfekt prositiven ($r = 1$) oder negativen ($r = -1$) Zusammenhang:

$$\begin{aligned} R^2 &= 1 - \frac{\sum e^2}{\sum (y_i - \bar{y})^2} \\ 1 - R^2 &= \frac{\sum e^2}{\sum (y_i - \bar{y})^2} \\ (1 - R^2) \sum (y_i - \bar{y})^2 &= \sum e^2 \end{aligned} \tag{1}$$

Wir wählen deshalb die erste Regression mit dem höchsten Determinationskoeffizient:

$$\begin{aligned} R_1^2 &= r_1^2 = 0.75^2 = 0.56 \\ R_2^2 &= r_2^2 = 0.49^2 = 0.24 \\ R_3^2 &= r_3^2 = 0.62^2 = 0.38 \end{aligned} \tag{2}$$

2. Wie gross ist die Gesamtkorrelation? Hinweis: Korrelationskoeffizienten können nach *Fishers z-Transformation* $z = \frac{1}{2}(ln(1 + r) - ln(1 - r))$ gemittelt werden.

$$\begin{aligned} z_1 &= \frac{1}{2}(ln(1 + r_1) - ln(1 - r_1)) = \frac{1}{2}(ln(1 + 0.75) - ln(1 - 0.75)) = 0.97 \\ z_2 &= \frac{1}{2}(ln(1 + r_2) - ln(1 - r_2)) = \frac{1}{2}(ln(1 + 0.49) - ln(1 - 0.49)) = 0.54 \\ z_2 &= \frac{1}{2}(ln(1 + r_3) - ln(1 - r_3)) = \frac{1}{2}(ln(1 + 0.62) - ln(1 - 0.62)) = 0.73 \\ \bar{z} &= \frac{\sum z_i}{n} = \frac{0.97 + 0.54 + .73}{3} = 0.75 \end{aligned}$$

Die Umkehrfunktion der Transformation ist der *Tangens hyperbolicus*:

$$\begin{aligned} z &= \frac{1}{2}(ln(1 + r) - ln(1 - r)) \\ 2z &= ln(1 + r) - ln(1 - r) \\ e^{2z} &= e^{ln(1+r) - ln(1-r)} = \frac{e^{ln(1+r)}}{e^{ln(1-r)}} = \frac{1 + r}{1 - r} \\ (1 - r)e^{2z} &= 1 + r \\ e^{2z} - e^{2z} r &= 1 + r \\ e^{2z} - 1 &= r + e^{2z} r \\ e^{2z} - 1 &= r(e^{2z} + 1) \\ r &= \frac{e^{2z} - 1}{e^{2z} + 1} = tanh(z) \end{aligned}$$

Die Gesamtkorrelation ist $\bar{r} = tanh(\bar{z}) = tanh(0.75) = 0.64$

Übung 5. *Regression*
Sie haben die Abzahl Ebola-Neuansteckungen über einen Zeitraum in einer bestimmten Region erhoben:

x_i	y_i
1	5
2	7
3	20
4	34
5	45

Wie hoch schätzen Sie die Inzidenz für x_{30}?

Der Scatterplot deutet auf linearen oder exponentiellen Zusammenhang hin.

x_i	x_i^2	y_i	$x_i y_i$	y_i'	$x_i y_i'$
1	1	5	5	1.61	1.61
2	4	7	14	1.94	3.90
3	9	20	60	3.00	8.99
4	16	34	136	3.53	14.10
5	25	45	22	3.80	19.03
15	55	111	440	13.88	47.63

1. Für das lineare Modell erhalten wir:

$$
\begin{aligned}
n &= 5 \\
\bar{x} &= \frac{\sum_1^n x_i}{n} = \frac{15}{5} = 3 \\
\bar{y} &= \frac{\sum_1^n y_i}{n} = \frac{111}{5} = 22.2 \\
b &= \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{440 - 5 \cdot 3 \cdot 22.2}{55 - 5 \cdot 3^2} = 10.7 \\
a &= \bar{y} - b\bar{x} = 22.2 - 10.7 \cdot 3 = -9.9
\end{aligned}
$$

   Damit ist $f(x) = a + bx = -9.9 + 10.7x$ und $f(10) = 97$.

2. Für das exponentielle Modell erhalten wir:

$$
\begin{aligned}
n &= 5 \\
\bar{x} &= \frac{\sum_1^n x_i}{n} = \frac{15}{5} = 3 \\
\bar{y} &= \frac{\sum_1^n y_i}{n} = \frac{13.88}{5} = 2.78 \\
b &= \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{47.63 - 5 \cdot 3 \cdot 2.78}{55 - 5 \cdot 3^2} =?? \\
a &= \bar{y} - b\bar{x} = 2.78 - 0.59 \cdot 3 = 1.01
\end{aligned}
$$

Damit ist $f(x) = e^{a+bx} = e^{1.01+0.59}x$ und $f(10) = 1002$.

Nach zehn Tagen stecken sich damit im besseren Fall rund 100 Personen pro Tag neu an, im schlechteren Fall rund 1000.