Fachbereich Mathematik
Prof. Dr. Michael Kohler
Dominique Küpper

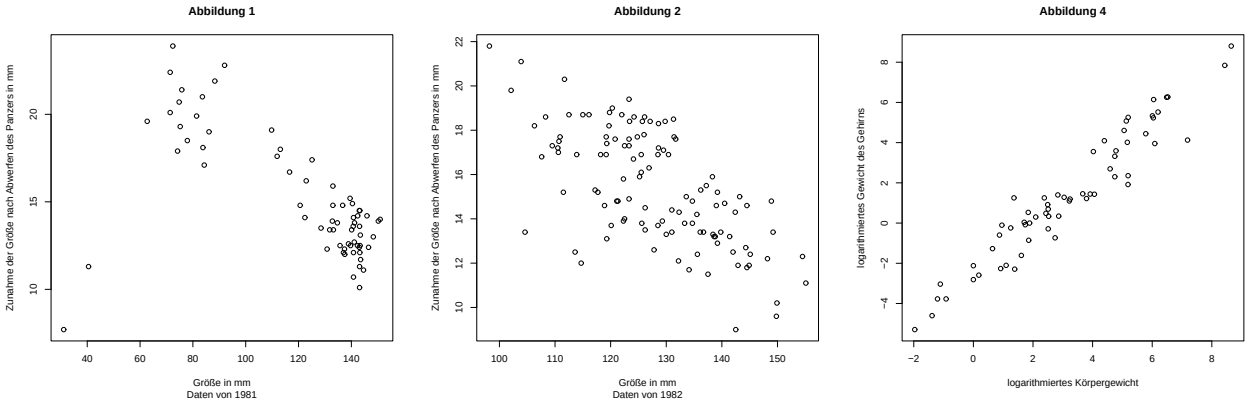
TECHNISCHE UNIVERSITÄT DARMSTADT

WS 09/10
24.11.2009

4. Übungsblatt zur „Mathematik und Statistik für Biologie“

Aufgabe 13 (Korrelation) (3 Punkte)

Gegeben seien die folgenden zwei Mengen von Datenpunkten:



Abbildungen 1 & 2 stellt den Zusammenhang zwischen der Größe eines Krebses und seiner Gewichtszunahme nach Abwerfen des Panzers dar. In Abbildung 3 ist der Zusammenhang zwischen dem logarithmierten Körpergewicht von Landsäugetieren und dem logarithmierten Gewicht ihres Gehirns abgebildet.

Welche Aussage können Sie über die Größe der Korrelation der Datenmengen machen (z.B $r_{x,y} = -1$, $-1 < r_{x,y} < 0$, $r_{x,y} = 0$, $0 < r_{x,y} < 1$ oder $r_{x,y} = 1$) ? Begründen Sie Ihre Aussage!

Lösung:

Abb 1+2 Aufgrund der Lage der Datenpunkte wird die Steigung der Regressionsgerade negativ sein. Da die Korrelation das gleiche Vorzeichen hat wie die Steigung und immer im Intervall $[-1, 1]$ liegt, folgt $-1 < r_{x,y} < 0$. (Da die Punkte nicht alle auf einer Geraden liegen, ist $r_{x,y} \neq -1$. Da die Regressionsgerade nicht waagerecht verläuft, ist auch $r_{x,y} \neq 0$.)

Abb 3 Da die Steigung der Regressionsgerade positiv ist, ist auch die empirische Korrelation positiv. Wie zuvor können 1 und 0 nicht vorkommen. Also gilt $r_{x,y} \in (0, 1)$.

Aufgabe 14 (2 Punkte)

In der folgenden Tabelle ist der Schuldenstand der Länder und Gemeinden je Einwohner in den einzelnen Bundesländern am 31.12.2008 aufgelistet (Quelle: Statistische Bundesamt):

Bundesland	**Schulden (Euro)**	**Bundesland**	**Schulden (Euro)**
Baden-Württemberg	4439	Niedersachsen	7218
Bayern	2861	Nordrhein-Westfalen	7620
Berlin	16340	Rheinland-Pfalz	7904
Brandenburg	7408	Saarland	10182
Bremen	23084	Sachsen	3229
Hamburg	12223	Sachsen-Anhalt	9467
Hessen	6344	Schleswig-Holstein	8677
Mecklenburg-Vorpommern	6893	Thüringen	7803

(a) Bestimmen Sie das empirische arithmetische Mittel.

(b) Warum stimmt es nicht mit der bundesweiten Verschuldung je Einwohner von 5866 Euro überein (die Schulden des Bundes sind auch hier nicht mitgerechnet)?

Lösung: Das empirische arithmetische Mittel $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 8856$ Euro stimmt nicht mit der bundesweiten Verschuldung je Einwohner von 5866 Euro überein, da dort die unterschiedlichen Bevölkerungszahlen der einzelnen Bundesländer eine Rolle spielen.

Aufgabe 15 (Lineare Regression) (3 Punkte)

Die folgende Tabelle enthält das durchschnittliche Gewicht von einigen Landsäugetieren und das mittlere Gewicht ihres Gehirns. Es gibt Untersuchungen, die einen linearen Zusammenhang zwischen den Logarithmen dieser beiden Größen sehen.

Name	Körpergewicht [kg]	Gewicht des Gehirns [g]
Kuh	465	423
Katze	3,3	25,6
Asiatischer Elefant	2547	4603

Bestimmen Sie die Regressionsgerade bzgl. der *logarithmierten* Datenpaare. Schätzen Sie mit Hilfe der von Ihnen berechneten Regressionsgeraden das Gewicht des Gehirns eines Gorillas mit einem Körpergewicht von 207 kg? (*Hinweis:* Laut Vorlesung ist die Formel für die Regressionsgerade: $y = \hat{a} \cdot (x - \bar{x}) + \bar{y}$ mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ und $\hat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.)

Lösung: Die Regressionsgerade soll zu den Daten

$$x_1 = \ln(465), y_1 = \ln(423), x_2 = \ln(3,3), y_2 = \ln(25,6), x_3 = \ln(2547), y_3 = \ln(4603)$$

bestimmt werden.
Die Regressionsgerade hat (vgl. Vorlesung) die Form:

$$y = \hat{a} \cdot (x - \bar{x}) + \bar{y}.$$

Im vorliegenden Fall bedeutet das:

$$\begin{aligned}
\bar{x} &= \frac{1}{3}\left(\ln(465) + \ln(3,3) + \ln(2547)\right) = \frac{1}{3} \cdot 15,17863135 = 5,059543783 \\
\bar{y} &= 5,908142691 \\
s_x^2 &= \frac{1}{2}\left((\ln(465) - \bar{x})^2 + (\ln(3,3) - \bar{x})^2 + (\ln(2547) - \bar{x})^2\right) = 11,93031017 \\
s_{x,y} &= \frac{1}{2}\left((\ln(465) - \bar{x})(\ln(423) - \bar{y}) + (\ln(3,3) - \bar{x})(\ln(25,6) - \bar{y}) + (\ln(2547) - \bar{x})(\ln(4603) - \bar{y})\right) \\
&= 8,742898383
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \quad \hat{a} = \frac{s_{x,y}}{s_x^2} = \frac{8,742898383}{11,93031017} = 0,7328307696 \\
&\Rightarrow \quad y = 0,73283 \cdot x + 2,20035
\end{aligned}$$

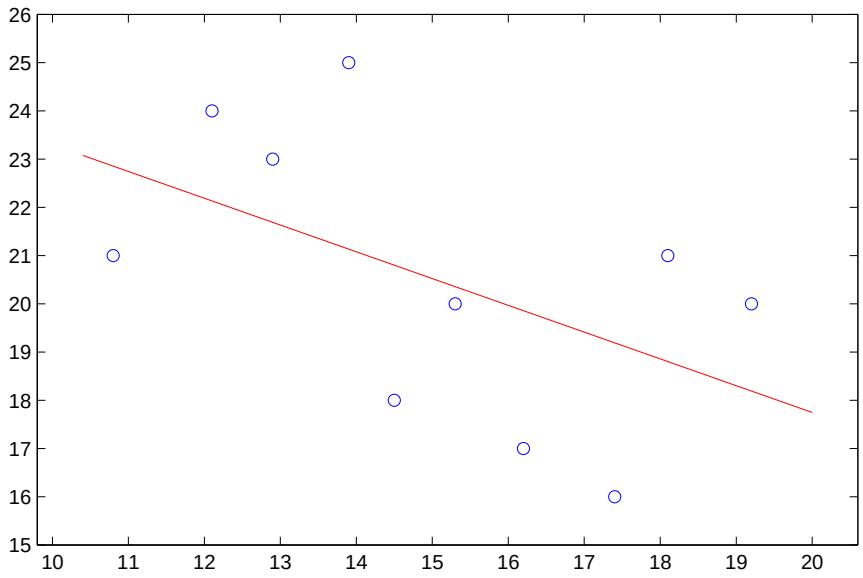
Setzt man für x jetzt $\ln(207)$ ein, so erhält man den Logarithmus vom Gewicht des Gehirns. Nach Anwenden der Exponentialfunktion erhält man, dass das Gehirn eines Gorillas etwa $449,5856$ Gramm wiegen müsste. (Der wirkliche Wert liegt bei 406 Gramm).

Aufgabe 16 (Lokale Mittelung) (4 Punkte)

In einer Fertigungsanlage kann eine der Maschinen durch eine Stellschraube justiert werden. Die Anzahl der Produktionsfehler lässt sich durch diese Schraube beeinflussen. Bei der Feinabstimmung wurden die folgenden Zahlen in Abhängigkeit von der Tiefe der Schraube beobachtet:

Tiefe (μm)	10,8	12,1	12,9	13,9	14,5	15,3	16,2	17,4	18,1	19,2
Fehlerzahl	21	24	23	25	18	20	17	16	21	20

Diese Daten sind in folgendem Scatterplot dargestellt, in dem auch schon die zugehörige Regressionsgerade eingezeichnet ist:



(a) Wir wollen nun eine Schätzung für die Fehlerzahlen bei den Tiefen $x = 11, x = 12, x = 13, \ldots, x = 20$ mittels *lokaler Mittelung* bestimmen. Berechnen Sie dazu das (arithmetische) Mittel aller Punkte, deren Abstand vom jeweils betrachteten x-Wert kleiner als die Schranke **h=1** entfernt ist und tragen Sie die Werte in folgende Tabelle ein.

x**-Wert (Tiefe)**	11	12	13	14	15	16	17	18	19	20
y**-Wert** ($h = 1$)										

(b) Tragen Sie alle in (a) berechneten Punkte in den Scatterplot ein und verbinden Sie diese.

(c) Vergleichen Sie das Ergebnis dieser *nichtparametrischen Regressionsschätzung* mit dem der linearen Regression (im Scatterplot).

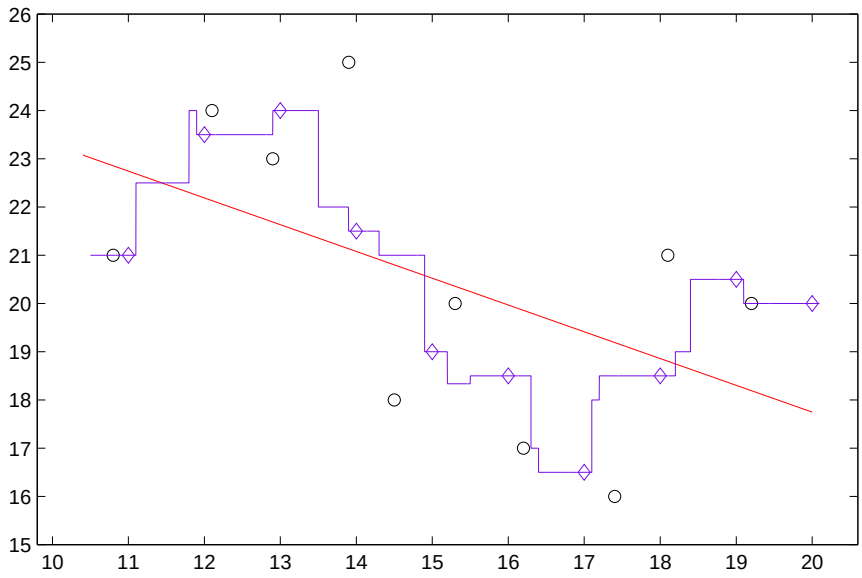
Lösung:

(a) Wir illustrieren das Prinzip an einem Beispiel: Zum x-Wert 14 haben die Messwerte $13,9$ und $14,5$ einen Abstand kleiner als 1. Damit ist der zugehörige y-Wert $\frac{1}{2}(25 + 18) = 21,5$.

Es ergeben sich die folgenden Werte:

x**-Wert (Tiefe)**	11	12	13	14	15	16	17	18	19	20
y**-Wert** ($h = 1$)	21	23,5	24	21,5	19	18,5	16,5	18,5	20,5	20

(b) Damit wird der Scatterplot zu:



(c) Im gegebenen Beispiel führt die Annahme, dass es einen linearen Zusammenhang gibt, auf die Vermutung, dass die Zahl der Produktionsfehler mit der Tiefe der Stellschraube immer weiter abnimmt. Diese Annahme muss aber nicht zutreffen. Es könnte auch ein nicht-linearer Zusammenhang bestehen.

Die Schätzung durch lokale Mittelung erlaubt es nicht-lineare Zusammenhänge zwischen den Daten zu erkennen. So kann man in unserem Beispiel vermuten, dass ein Minimum der Fehlerzahlen in der Produktion zwischen $16\mu m$ und $17\mu m$ erreicht wird. Die geringe Zahl der Messungen lassen allerdings keine gesicherten Aussagen zu - die beiden höheren Werte am Ende, die bei der lokalen Mittelung zu einem Anstieg der geschätzten Produktionsfehler bei größeren Tiefen führen, könnten auch durch weitere Einflüsse entstanden sein.

Außerdem ist bei diesem Verfahren die Wahl der Schranke des x-Abstands h von großer Bedeutung. Ist diese zu klein gewählt, mitteln sich Messfehler nicht mehr genügend heraus. Die Schätzung spiegelt dann zwar die gegebene Messreihe sehr genau wieder, aber nicht unbedingt den zu ermittelnden Zusammenhang.