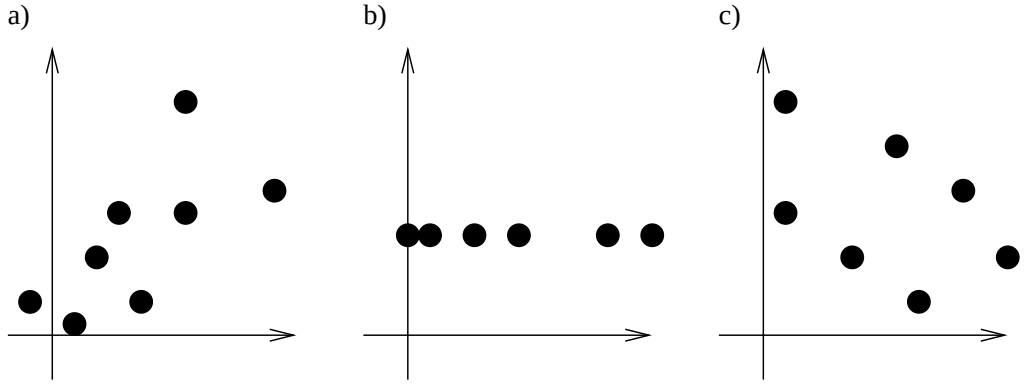
Fachbereich Mathematik
Prof. Dr. Michael Kohler
Andreas Gärtner
Jens Mehnert

TECHNISCHE UNIVERSITÄT DARMSTADT

WS 09/10
25./26.11.2009

4. Übungsblatt zur „Statistik I für Human- und Sozialwissenschaft“

Aufgabe 13 (Korrelation) (3 Punkte)



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:

(a) Durch die erste Datenmenge kann man eine steigende Regressionsgerade legen. Daher gilt für die Korrelation $0 < r_{x,y} < 1$.

(b) Durch Datenmenge b) lässt sich eine Gerade legen die alle Datenpunkte enthält. Daher gilt $r_{x,y} = 0$.

(c) Durch Datenmenge c) läßt sich eine fallenden Gerade legen. Daher gilt $-1 < r_{x,y} < 0$.

Aufgabe 14 (Lineare Regression) (3 Punkte)

Fünf Versuchspersonen absolvieren zwei unterschiedliche psychologische Tests zur Kraftfahreignung. Dabei entstehen die folgenden Testergebnisse:

Versuchsperson-Nr.	Test x	Test y
1	31	15
2	128	95
3	67	35
4	46	40
5	180	80

(a) Zeichnen Sie die Punkte (Testergebnis in Test x, Testergebnis in Test y) in ein Streudiagramm (Scatterplot).

(b) Berechnen Sie die zugehörige Regressionsgerade und zeichnen Sie diese in das Diagramm ein. (*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 $\hat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.)

(c) Interpretieren Sie den Verlauf der Regressionsgerade bzgl. eines Zusammenhanges zwischen Test x und Test y.

Lösung:

(b) Wir rechnen mit obigen Formeln aus $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 90,4$, $\bar{y} = 53$, $\hat{a} = \frac{7254}{15489,2} = 0,468326318 \approx 0,47$ und erhalten damit $y = 0,47 \cdot (x - 90,4) + 53$.

(c) Die Regressionsgerade ist stark steigend. D.h. je höher das Testergebnis einer Versuchsperson in Test x ist umso höher ist im Mittel das Testergebnis derselben Person auch in Test y. Dies spricht für einen Zusammenhang zwischen beiden Tests.

Aufgabe 15 (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 auch hier nicht mitgerechnet)?

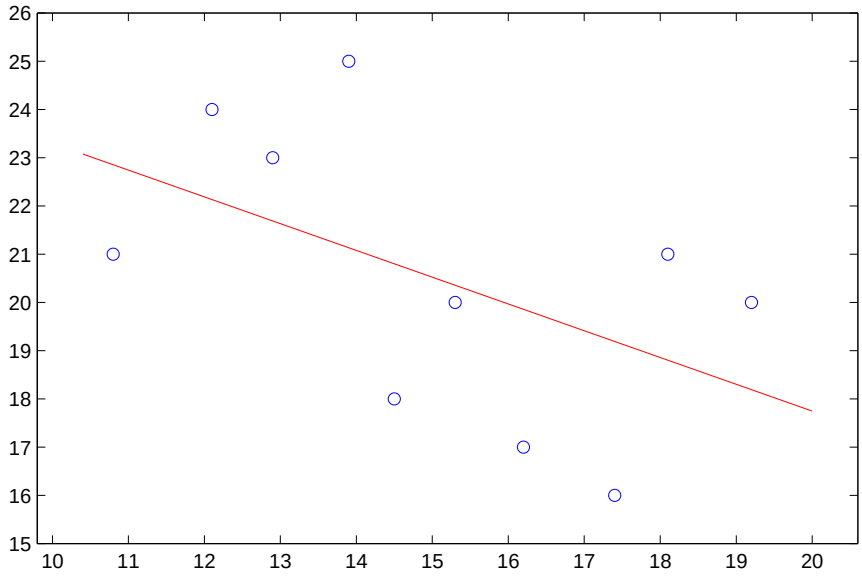
Lösung: Das empirische arithmetische Mittel $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 8856$ Euro stimmt nicht mit der bundesweiten Verschuldung je Einwohner von 5866 Euro überein, da dort die unterschiedlichen Bevölkerungszahlen der einzelnen Bundesländer eine Rolle spielen.

Aufgabe 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.

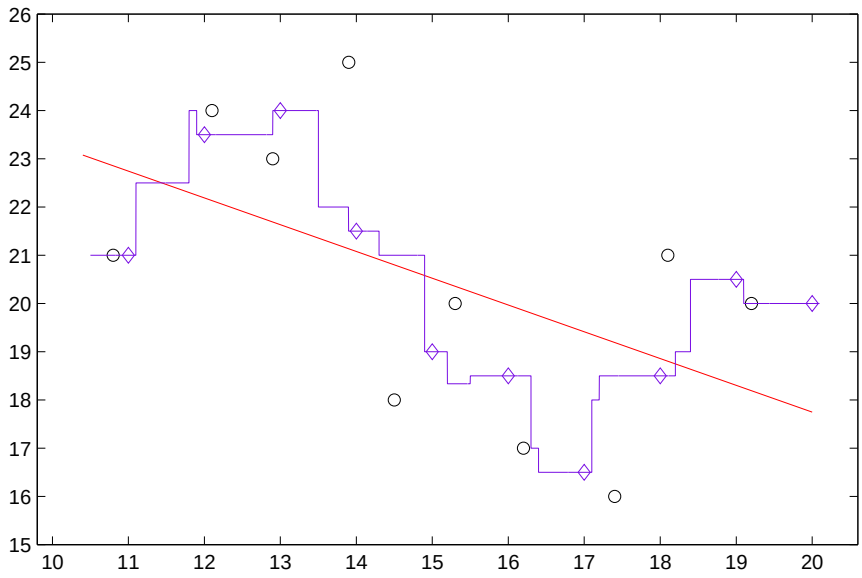
(c) Vergleichen Sie das Ergebnis dieser *nichtparametrischen Regressionsschätzung* mit dem der linearen Regression (im Scatterplot).

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.