Einführung in die Stochastik
3. Übungsblatt

TECHNISCHE UNIVERSITÄT DARMSTADT

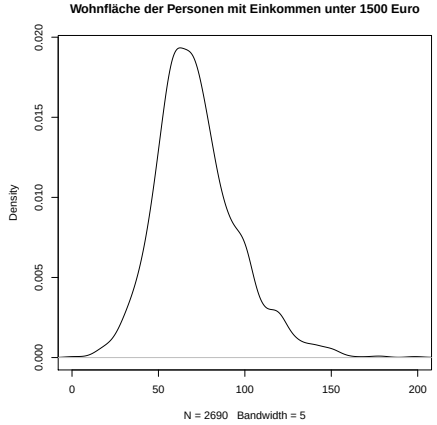
Fachbereich Mathematik
M. Kohler
A. Fromkorth
D. Furer

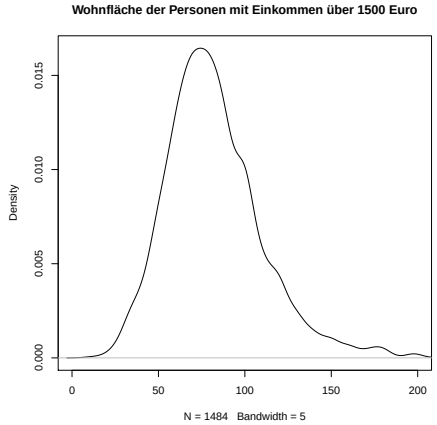
SS 2011
06.05.2011

Gruppen und Hausübung

Aufgabe 9 (4 Punkte)

Der Mikrozensus ist eine statistische Erhebung. Hierbei werden nach bestimmten Zufallskriterien Haushalte ausgewählt, die Daten zu unteschiedlichen Merkmalen liefern, wie z.B. Nettoeinkommen, Alter, Arbeitszeit, Wohnfläche. In den unten stehenden Abbildungen sind Kern-Dichteschätzer (mit Gauß-Kern) angewandt auf die Wohnfläche bei Personen mit einem Nettoeinkommen unter 1500 Euro und mit einem Nettoeinkommen von über 1500 Euro. Welche Aussagen lassen sich anhand dieser Grafiken treffen? Begründen Sie diese.





Hinweis: Betrachten Sie z.B. als Referenz die Stelle 100 m^2.

Lösung: Vergleicht man diese untereinander liegende Dichteschätzer, so stellt man eine Verlagerung nach rechts bei einem Nettoeinkommen von über 1500 Euro fest. Dies bedeutet nichts Anderes als, dass der relative Anteil der Personen die sich eine größere Wohnung leisten mit dem Einkommen steigt. Diesen Zusammenhang stellt man fest, wenn man z.B. die Flächen unter den Grafen auf dem Abschnitt $[100, 200]$ vergleicht.

Aufgabe 10 (4 Punkte)

In der folgenden Tabelle sind die Ausgaben pro Student (in Euro) und die Arbeitslosenquote (in Prozent) in den sechs neuen Bundesländern im Jahr 2001 angegeben.

	Ausgaben pro Student (in Euro)	Arbeitslosenquote (in Prozent)
Berlin	8100	17.9
Brandenburg	6600	18.8
M.-V.	8700	19.6
Sachsen	8700	19
Sachsen-Anhalt	9900	20.9
Thüringen	8800	16.5

(a) Zeichnen Sie ein Streudiagramm (Scatterplot) der Daten, wobei sie als x-Wert die Ausgaben pro Student und als y-Wert die Arbeitslosenquote verwenden.

(b) Bestimmen Sie mit Hilfe der in der Vorlesung hergeleiteten allgemeinen Formel die zugehörige Regressionsgerade und zeichnen Sie diese in das Streudiagramm aus a) ein.

(c) Inwieweit ändert sich das Resultat in b), wenn man den zu Sachsen-Anhalt gehörenden Datenpunkt weglässt?

Lösung:

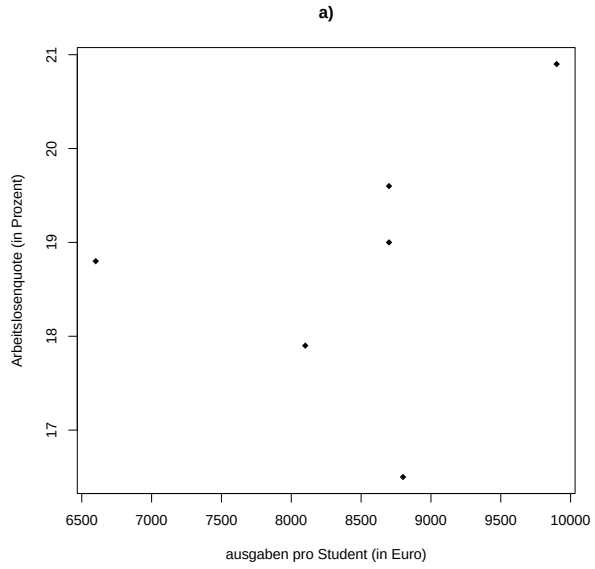


Abbildung 1: Aufgabe 10 a)

(b) Seien x_i die Ausgaben pro Student (in Euro) in Zeile i der Tabelle und y_i die Arbeitslosenquote (in Prozent) in Zeile i der Tabelle. Nach der Formel aus der Vorlesung hat die Regressionsgerade die Form

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

mit

$$\begin{aligned}
\hat{a} &= \frac{s_{xy}}{s_x^2}, \\
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}), \\
s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \\
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n}x_i, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n}y_i
\end{aligned}$$

und $n = 6$. Einsetzen der Werte ergibt

$$\begin{aligned}
\bar{x} &= \frac{1}{6}(8100+6600+8700+8700+9900+8800) \approx 8466.667 \\
\bar{y} &= \frac{1}{6}(17.9+18.8+19.6+19+20.9+16.5) \approx 18.7833 \\
s_{xy} &= \frac{1}{5}((8100-8466.667)\cdot(17.9-18.7833)+\cdots+(8800-8466.667)\cdot(16.5-18.7833)) \\
&\approx 561.33 \\
s_x^2 &= \frac{1}{5}((8100-8466.667)^2+\cdots+(8800-8466.667)^2) \approx 1178667 \\
\hat{a} &\approx 0.000476
\end{aligned}$$

und damit

$$y = 0.000476\cdot(x-8466.667)+18.7833.$$

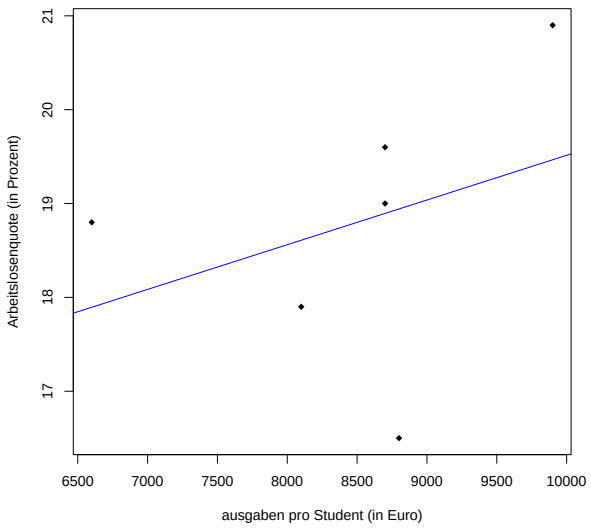


Abbildung 2: Aufgabe 10 b)

(c) Lässt man den Sachsen-Anhalt Datenpunkt weg, so wird die Steigung der Regressionsgeraden negativ (was auch zu einer Änderung des y-Achsenabschnitts führt).

$$\begin{aligned} \bar{x} &= 8180 \\ \bar{y} &= 18.36 \\ \hat{a} &= -0.0002432905 \end{aligned}$$

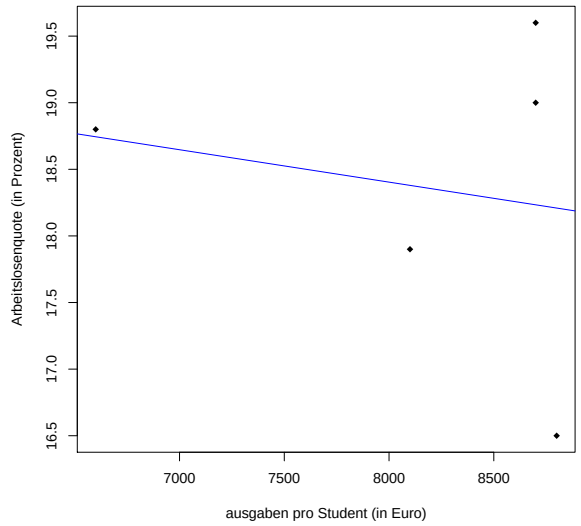


Abbildung 3: Aufgabe 10 c)

Aufgabe 11 (4 Punkte)

(a) Seien $x_1, y_1, \ldots, x_n, y_n \in \mathbb{R}$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Zeigen Sie:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$$

und

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

(b) Berechnen Sie die Korrelation der Daten aus Aufgabe 10.

(c) Was folgt aus b) für die Steigung der zugehörigen Regressionsgeraden ?

(d) Inwieweit ändert sich das Ergebnis aus b), wenn man vor Beginn der Berechnung der Korrelation die Ausgaben pro Student in Dollar und die Arbeitslosenquote in Promille umrechnet? Begründen Sie ihre Antwort.

Lösung:

(a)

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\\
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \frac{1}{n}\sum_{i=1}^{n}\left(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i-\bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y}\\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}
\end{aligned}$$

(b) Die empirische Korrelation ist definiert als

$$r_{xy}=\frac{s_{xy}}{s_x\cdot s_y}.$$

Die Werte von s_{xy} und s_x wurden schon in Aufgabe 10 berechnet. Wegen

$$s_y^2\approx 2.237667$$

folgt dann

$$r_{xy}=\frac{s_{xy}}{\sqrt{s_x^2\cdot s_y^2}}\approx 0.34564.$$

(c) Da das Vorzeichen der empirischen Korrelation mit dem Vorzeichen der Steigung der Regressionsgeraden übereinstimmt, ist im vorliegenden Fall die Steigung der Regressionsgeraden positiv, da $r_{xy}=0.3456429>0$.

(d) Das Umrechnen der Einheiten kann man als Multiplikation mit einer positiven Konstanten realisieren. Anstelle der Daten $(x_1,y_1),\ldots,(x_n,y_n)$ betrachten wir also die Datenpunkte $(z_1,w_1),\ldots,(z_n,w_n)$ mit $(z_i,w_i)=(ax_i,by_i)$ für $i=1,\ldots,n$. Dann gilt:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n}z_i=\frac{1}{n}\sum_{i=1}^{n}ax_i=a\frac{1}{n}\sum_{i=1}^{n}x_i=a\cdot\bar{x}$$

und genauso

$$\bar{w} = b\cdot\bar{y}.$$

Damit erhalten wir

$$\begin{aligned}
s_{zw} &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})\cdot(w_i-\bar{w}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})\cdot(by_i-b\bar{y}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}a(x_i-\bar{x})\cdot b(y_i-\bar{y}) \\
&= ab s_{xy}, \\
s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}a^2(x_i-\bar{x})^2 \\
&= a^2 s_x^2.
\end{aligned}$$

und

$$\begin{aligned}
s_w^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(w_i-\bar{w})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(by_i-b\bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}b^2(y_i-\bar{y})^2 \\
&= b^2 s_y^2.
\end{aligned}$$

Das bedeutet für die empirische Korrelation

$$r_{zw} = \frac{s_{zw}}{\sqrt{s_z^2 s_w^2}} = \frac{ab s_{xy}}{\sqrt{a^2 s_x^2 b^2 s_y^2}} = r_{xy},$$

d.h. die empirsiche Korrelation ändert sich durch die Umrechnung nicht.

Aufgabe 12 (4 Punkte)

Gegeben sei eine zweidimensionale Messreihe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

vom Umfang n. Anstelle einer Geraden (wie bei der linearen Regression) könnte man analog auch einen Polynom dritten Grades

$$y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

durch Minimierung von

$$F(a,b,c,d) := \sum_{i=1}^{n}\left(y_i - (a + b \cdot x_i + c \cdot x_i^2 + d \cdot x_i^3)\right)^2$$

an die Daten anpassen. Zeigen Sie (durch Nullsetzen geeigneter Ableitungen), dass die Werte a, b, c, d, für die $F(a,b,c,d)$ minimal wird, Lösungen des linearen Gleichungssystems

$$\begin{aligned}
a + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + c\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 + d\frac{1}{n}\sum_{i=1}^{n} x_i^3 &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 + c\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^3 + d\frac{1}{n}\sum_{i=1}^{n} x_i^4 &= \frac{1}{n}\sum_{i=1}^{n} x_i\cdot y_i \\
a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^2 + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^3 + c\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^4 + d\frac{1}{n}\sum_{i=1}^{n} x_i^5 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2\cdot y_i \\
a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^3 + b\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^4 + c\cdot\frac{1}{n}\sum_{i=1}^{n} x_i^5 + d\frac{1}{n}\sum_{i=1}^{n} x_i^6 &= \frac{1}{n}\sum_{i=1}^{n} x_i^3\cdot y_i
\end{aligned}$$

sind.

Lösung: Wie in der Vorlesung müssen wir die partiellen Ableitungen nullsetzen. Dies ergibt

$$\frac{\partial}{\partial a}F(a,b,c,d) = \frac{\partial}{\partial b}F(a,b,c,d) = \frac{\partial}{\partial c}F(a,b,c,d) = \frac{\partial}{\partial d}F(a,b,c,d) = 0.$$

Wir berechnen also die partiellen Ableitungen

$$\begin{aligned}
0 = \frac{\partial}{\partial a}F(a,b,c,d) &= \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}\left((y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) = \sum_{i=1}^{n}\frac{\partial}{\partial a}\left(\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) \\
&= \sum_{i=1}^{n} 2\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\frac{\partial}{\partial a}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right) \\
&= -2\sum_{i=1}^{n}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right) \\
&= -2\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i - \sum_{i=1}^{n} cx_i^2 - \sum_{i=1}^{n} dx_i^3\right) \\
\Rightarrow \quad & \frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{1}{n}\sum_{i=1}^{n} x_i + c\frac{1}{n}\sum_{i=1}^{n} x_i^2 + d\frac{1}{n}\sum_{i=1}^{n} x_i^3.
\end{aligned}$$

$$\begin{aligned}
0 = \frac{\partial}{\partial b}F(a,b,c,d) &= \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}\left((y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) = \sum_{i=1}^{n}\frac{\partial}{\partial b}\left(\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) \\
&= \sum_{i=1}^{n} 2\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\frac{\partial}{\partial b}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right) \\
&= -2\sum_{i=1}^{n}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\cdot x_i \\
&= -2\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} bx_i^2 - \sum_{i=1}^{n} cx_i^3 - \sum_{i=1}^{n} dx_i^4\right) \\
\Rightarrow \quad & \frac{1}{n}\sum_{i=1}^{n} x_i y_i = a\frac{1}{n}\sum_{i=1}^{n} x_i + b\frac{1}{n}\sum_{i=1}^{n} x_i^2 + c\frac{1}{n}\sum_{i=1}^{n} x_i^3 + d\frac{1}{n}\sum_{i=1}^{n} x_i^4.
\end{aligned}$$

und

$$\begin{aligned}
0 = \frac{\partial}{\partial c} F(a,b,c,d) &= \frac{\partial}{\partial c}\left(\sum_{i=1}^{n}\left((y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) = \sum_{i=1}^{n}\frac{\partial}{\partial c}\left(\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) \\
&= \sum_{i=1}^{n} 2\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\frac{\partial}{\partial c}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right) \\
&= -2\sum_{i=1}^{n}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\cdot x_i^2 \\
&= -2\left(\sum_{i=1}^{n} x_i^2 y_i - \sum_{i=1}^{n} a x_i^2 - \sum_{i=1}^{n} b x_i^3 - \sum_{i=1}^{n} c x_i^4 - \sum_{i=1}^{n} d x_i^5\right) \\
\Rightarrow \quad & \frac{1}{n}\sum_{i=1}^{n} x_i^2 y_i = a\frac{1}{n}\sum_{i=1}^{n} x_i^2 + b\frac{1}{n}\sum_{i=1}^{n} x_i^3 + c\frac{1}{n}\sum_{i=1}^{n} x_i^4 + d\frac{1}{n}\sum_{i=1}^{n} x_i^5.
\end{aligned}$$

und

$$\begin{aligned}
0 = \frac{\partial}{\partial d} F(a,b,c,d) &= \frac{\partial}{\partial d}\left(\sum_{i=1}^{n}\left((y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) = \sum_{i=1}^{n}\frac{\partial}{\partial d}\left(\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)^2\right) \\
&= \sum_{i=1}^{n} 2\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\frac{\partial}{\partial d}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right) \\
&= -2\sum_{i=1}^{n}\left(y_i - (a + bx_i + cx_i^2 + dx_i^3)\right)\cdot x_i^3 \\
&= -2\left(\sum_{i=1}^{n} x_i^3 y_i - \sum_{i=1}^{n} a x_i^3 - \sum_{i=1}^{n} b x_i^4 - \sum_{i=1}^{n} c x_i^5 - \sum_{i=1}^{n} d x_i^6\right) \\
\Rightarrow \quad & \frac{1}{n}\sum_{i=1}^{n} x_i^3 y_i = a\frac{1}{n}\sum_{i=1}^{n} x_i^3 + b\frac{1}{n}\sum_{i=1}^{n} x_i^4 + c\frac{1}{n}\sum_{i=1}^{n} x_i^5 + d\frac{1}{n}\sum_{i=1}^{n} x_i^6.
\end{aligned}$$

Dies sind die gewünschten Gleichungen.
Anmerkung für Studenten ab dem 3. Semester: Um zu zeigen, dass es sich in der Tat um ein Minimum handelt, muss man jetzt noch nachrechnen, dass die zugehörige Hessematrix positiv definit ist.

Dieses Übungsblatt wird im Rahmen der Übungen am 09. bzw. 10.05.2011 besprochen. Ihre Ausarbeitungen geben Sie am 16. bzw. 17.05.2011 in Ihrere Übungsgruppe ab. Sie erhalten diese am 23. bzw. 24.05.2011 korrigiert zurück.