**Universität Bremen**
**Fachbereich Mathematik**

Stochastik
Sommersemester 2007

# Lösungsskizzen zur 13. Übung

## Präzensübung Aufgabe 1

**b)** Um die Regressionsgerade zu berechnen brauchen wir die Mittelwerte $\overline{x}$, $\overline{y}$, die Steigung $\hat{m}$ und den y-Achsenabschnitt $\hat{b}$. Dabei ist

$$\begin{aligned} \overline{x} &= \tfrac{1}{n} \cdot \textstyle\sum_{i=1}^{n} x_i \\ \overline{y} &= \tfrac{1}{n} \cdot \textstyle\sum_{i=1}^{n} y_i \\ \hat{m} &= \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \hat{b} &= \overline{y} - \hat{m} \cdot \overline{x}. \end{aligned}$$

Aus den Werten der Tabelle können wir ausrechnen, dass

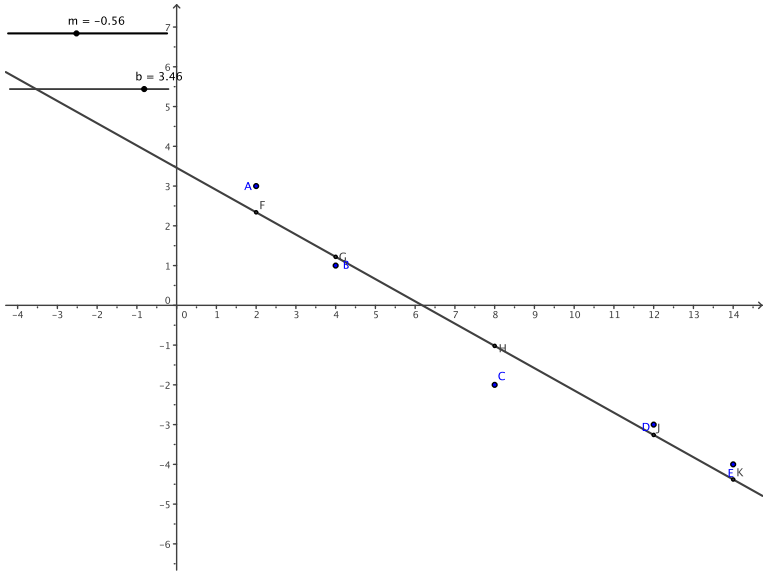
$$\overline{x} = 8 \text{ und } \overline{y} = -1.$$

Daraus ergibt sich die folgende Tabelle

| $x_i - \overline{x}$ | $-6$ | $-4$ | $0$ | $4$ | $6$ |
|---|---|---|---|---|---|
| $x_i$ | 2 | 4 | 8 | 12 | 14 |
| $y_i$ | 3 | 1 | $-2$ | $-3$ | $-4$ |
| $y_i - \overline{y}$ | 4 | 2 | $-1$ | $-2$ | $-3$ |

Setzen wir jetzt die Werte in die Formeln von $\hat{m}$ und $\hat{b}$ ein, so bekommen wir

$$\hat{m} = -\tfrac{29}{52} \approx -0,56 \text{ und } \hat{b} = -\tfrac{45}{13} \approx 3,46.$$




c) Die spannendere Frage als etwas auszurechnen ist, warum man die Steigung und den y-Achsenabschnitt der Regressionsgeraden auf diese Weise berechnen kann. Um sich dem zu nähern, machen wir den folgenden Ansatz:

1. Die Gerade soll durch den Punkt $(\overline{x}, \overline{y})$ gehen.
2. Der y-Achsenabschnitt soll so gewählt sein, dass die Summe der Abstände zwischen $(x_i, y_i)$ und $(x_i, y_i^*)$ minimal ist. Dabei ist $y_i^*$ gerade der $y$-Wert, wenn wir den $x_i$-Wert in die Geradengleichung eingeben. Im obigen Bild ist z.B. für den Wert $x_1 = 2$ der Punkt F gerade $(x_1, y_1^*)$.

Es ist nun nicht zwingend, wie wir die Abstände minimieren. Zunächst möchte man gerne die Summe der Beträge minimieren, also

$$\sum_{i=1}^{n} |y_i - y_i^*|.$$

Wenn man das Problem auf diese Weise modelliert, ist jedoch das Minimieren kompliziert. Man möchte meist auf Betragstriche verzichten. Um dennoch sicher zu gehen, dass man eine positive Zahl minimiert, ist nun der Trick, dass man die Quadrate betrachtet, also

$$\sum_{i=1}^{n} (y_i - y_i^*)^2.$$

Sehen wir uns diese Summe in Abhängigkeit von $\hat{m}$ an. Die Regressionsgerade hat die Funktionsgleichung

$$y = \hat{m}x + \hat{b}.$$

Da der Punkt $(\overline{x}, \overline{y})$ auf der Geraden liegt, gilt

$$\overline{y} = \hat{m}\overline{x} + \hat{b}.$$

Dies können wir umstellen zu

$$\hat{b} = \overline{y} - \hat{m}\overline{x}.$$

Das setzen wir ein in die Gleichung

$$y_i^* = \hat{m}x_i + \hat{b}$$

und bekommen so (nach geschicktem Klammern)

$$y_i^* = \hat{m}(x_i - \overline{x}) + \overline{y}.$$

Damit ist (durch Einsetzen von $y_i^*$)

$$S(m) \coloneqq \sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}(y_i - \overline{y} - m(x_i - \overline{x}))^2.$$

Wir erinnern uns, dass wir $S(m)$ minimieren wollen. Wir wollen also das $m$ herausfinden, so dass $S(m)$ minimal ist. Dieses $m$ nennen wir dann $\hat{m}$.

Man könnte nun $S(m)$ allgemein ableiten (siehe Exkurs am Ende), dann $S'(m) = 0$ setzen, sehen, dass $S''(m) > 0$ und so herausfinden, dass $S(m)$ genau dann minimal ist, wenn

$$\hat{m} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Statt dies allgemein zu tun, ist nun die Aufgabe, den Zusammenhang für die konkreten Werte der Aufgabe zu bestätigen.

Wenn wir die Werte einsetzen, erhalten wir

$$\begin{aligned} S(m) &= \sum_{i=1}^{5}(y_i - \overline{y} - m(x_i - \overline{x}))^2 \\ &= (4 + 6m)^2 + (2 + 4m)^2 + (-1)^2 + (-2 - 4m)^2 + (-3 - 6m)^2. \end{aligned}$$

An dieser Stelle können wir natürlich die Quadrate alle ausmultiplizieren, wir können aber auch mit der Kettenregel

$$[f(g(m))]' = f'(g(m)) \cdot g'(m)$$

sofort ableiten:

$$\begin{aligned} S'(m) &= 2 \cdot (4 + 6m) \cdot 6 + 2 \cdot (2 + 4m) \cdot 4 + 2 \cdot (-2 - 4m) \cdot (-4) + 2 \cdot (-3 - 6m) \cdot (-6) \\ &= 116 + 208m. \end{aligned}$$

Die zweite Ableitung ist:

$$S''(m) = 208$$

also immer größer 0. Daher bekommen wir durch das Nullsetzen von $S'(m)$ einen Minimalpunkt heraus. Wenn wir nun verlangen, dass

$$S'(m) = 0$$

dann ist

$$m = -\frac{116}{208} = -\frac{29}{52} \approx -0,56.$$

Dies ist exakt der selbe Wert wie der, den wir in Aufgabe b) für $\hat{m}$ herausbekommen haben.

**Exkurs** Da das allgemeine Ableiten der Funktionsgleichung

$$S(m) = \sum_{i=1}^{n}(y_i - \overline{y} - m(x_i - \overline{x}))^2$$

nicht lange dauert, führe ich es hier für die, die es interessiert, noch auf.

Wir leiten wieder mit Hilfe der Kettenregel ab. Die Schwierigkeit zu vorher besteht darin, dass wir nun die Variablen $x_i, y_i, \overline{x}$ und $\overline{y}$ als Konstanten betrachten und wie normale Zahlen beim Ableiten behandeln müssen. Wir bekommen

$$S'(m) = \sum_{i=1}^{n} 2 \cdot (y_i - \overline{y} - m \cdot (x_i - \overline{x})) \cdot (-(x_i - \overline{x}))$$

und für die zweite Ableitung

$$S''(m) = \sum_{i=1}^{n} 2 \cdot (x_i - \overline{x})^2.$$

Da das Quadrat jeder reellen Zahl größer 0 ist, ist auch die zweite Ableitung für jedes $m$ größer 0. Wir erhalten also durch das Nullsetzen von $S'(m)$ einen Minimalpunkt. Es sei nun

$$S'(m) = 0.$$

Dann ist

$$\begin{aligned}
& 0 = \sum_{i=1}^{n} 2 \cdot (y_i - \overline{y} - m \cdot (x_i - \overline{x})) \cdot (-(x_i - \overline{x})) \\
\Leftrightarrow \quad & 0 = -2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + 2 \cdot m \cdot \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
\Leftrightarrow \quad & m \cdot \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) \\
\Leftrightarrow \quad & m = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.
\end{aligned}$$

Wir haben genau die Formel erhalten, die in Aufgabe 1 b) angewendet wurde, um $\hat{m}$ zu errechnen.