Name:_____ Datum:_____

Zur Bestimmung des Terms der Regressionsgeraden

- Ausgangspunkt unser Überlegungen ist ein bivariater Satz von Daten

$$((x_1; y_1); (x_2; y_2); ....; (x_n; y_n))$$

mit den entsprechenden Mittelwerten $\overline{x}$ und $\overline{y}$, den Varianzen V_y und V_x ,der Kovarianz c_{xy} und dem Korrelationskoeffizienten r.

- Gesucht sind die Werte des **Steigungsfaktors a** und des **Ordinatenabschnitts b** des **Terms der Regressionsgeraden**

$$y(x) = a \cdot x + b \,.$$

- Dabei ist der Begriff Regressionsgerade über folgende Bedingung festgelegt:

Die **Summe der Abweichungsquadrate** zwischen den **Messwerten y_i** und den entsprechenden **Funktionswerten der Geraden** $y(x_i) = a \cdot x_i + b$, also

$$\begin{aligned} Q(a; b) &= \sum_{i=1}^{n} \left(y_i - y(x_i)\right)^2 \\ &= (y_1 - y(x_1))^2 + ... + (y_n - y(x_n))^2 \\ &= \left((y_1 - a \cdot x_1 - b)^2 + ... + (y_n - a \cdot x_n - b)^2\right) \end{aligned}$$

soll minimal werden.

Beachte: Der obige Term enthält die zwei Variablen a und b; die übrigen „Formvariablen" x_i und y_i muss man sich mit den entsprechenden Messwerten belegt vorstellen.

- Die folgenden Umformungen des Terms Q(a;b) überführen diesen in eine andere Form, an der man erkennt, welche Werte a und b haben müssen, damit Q(a; b) einen minimalen Wert hat.
- Diese Umformungen sind etwas kompliziert. Das liegt daran, dass die Terme viele verschiedene Symbole enthalten und dass der Sinn der einzelnen Teilschritte nur dann verständlich ist, wenn man schon im Voraus weiß, wohin man eigentlich will. Somit besteht Ihre Aufgabe darin, die einzelnen Rechenschritte und die Überlegungen nachzuvollziehen.

1. Schritt

In jeder quadrierten Klammer der Summe Q(a; b) addiert man „eine **komplizierte Null**": komplizierte Nullen sind zum Beispiel „7-7" oder „ $\overline{y} - \overline{y}$ ", u.s.w. Da die Addition eines Terms vom Wert Null den Wert des Klammerterms und damit den Wert von Q(a; b) nicht verändert, ist diese Umformung erlaubt. Wir addieren in jeder Klammer

$$0 = \overline{y} - \overline{y} + a \cdot \overline{x} - a \cdot \overline{x}$$

und erhalten

$$\begin{aligned} Q(a; b) &= \sum_{i=1}^{n} \left((y_i - a \cdot x_i - b)\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - a \cdot x_i - b + \overline{y} - \overline{y} + a \cdot \overline{x} - a \cdot \overline{x}\right)^2 \\ &= \sum_{i=1}^{n} \left([y_i - \overline{y}] - a \cdot [x_i - \overline{x}] - b + \overline{y} - a \cdot \overline{x}\right)^2 \\ &= \sum_{i=1}^{n} \left([y_i - \overline{y}] - a \cdot [x_i - \overline{x}] + k\right)^2 \end{aligned}$$

Dabei wurde der Term $\overline{y} - a \cdot \overline{x} - b$ mit k abgekürzt.

2. Schritt

Nun wird jede Klammer innerhalb der großen Summe ausquadriert und man erhält

$$Q(a; b) = \sum_{i=1}^{n} \left([y_i - \overline{y}]^2 + a^2 \cdot [x_i - \overline{x}]^2 + k^2 - 2 \cdot a \cdot [y_i - \overline{y}] \cdot [x_i - \overline{x}] + 2 \cdot k \cdot [y_i - \overline{y}] - 2 \cdot a \cdot k \cdot [x_i - \overline{x}] \right)$$

Schreibt man die Summe aus, so sieht man den riesigen Lappen.

$$\begin{aligned} Q(a; b) = {} & [y_1 - \overline{y}]^2 + a^2 \cdot [x_1 - \overline{x}]^2 + k^2 - 2 \cdot a \cdot [y_1 - \overline{y}] \cdot [x_1 - \overline{x}] + 2 \cdot k \cdot [y_1 - \overline{y}] - 2 \cdot a \cdot k \cdot [x_1 - \overline{x}] \\ & + \\ & [y_2 - \overline{y}]^2 + a^2 \cdot [x_2 - \overline{x}]^2 + k^2 - 2 \cdot a \cdot [y_2 - \overline{y}] \cdot [x_2 - \overline{x}] + 2 \cdot k \cdot [y_2 - \overline{y}] - 2 \cdot a \cdot k \cdot [x_2 - \overline{x}] \\ & + \\ & \vdots \\ & + \\ & [y_n - \overline{y}]^2 + a^2 \cdot [x_n - \overline{x}]^2 + k^2 - 2 \cdot a \cdot [y_n - \overline{y}] \cdot [x_n - \overline{x}] + 2 \cdot k \cdot [y_n - \overline{y}] - 2 \cdot a \cdot k \cdot [x_n - \overline{x}] \end{aligned}$$

3. Schritt:

Wir ordnen nun diese große Summe um, indem wir zunächst jede einzelne Spalte addieren, also zunächst die Summanden addieren, die in der Form sich ähneln. Zur Verdeutlichung ist die entsprechende erste Summe oben und im Folgenden gekennzeichnet. Als Abkürzung benutzen wir wieder das Summenzeichen.

$$Q(a; b) = \sum_{i=1}^{n} [y_i - \overline{y}]^2 + \sum_{i=1}^{n} a^2 \cdot [x_i - \overline{x}]^2 + \sum_{i=1}^{n} (-2 \cdot a) [y_i - \overline{y}] \cdot [x_i - \overline{x}]$$

$$+ n \cdot k^2 + \sum_{i=1}^{n} 2 \cdot k \cdot [y_i - \overline{y}] + \sum_{i=1}^{n} (-2 \cdot k \cdot a) [x_i - \overline{x}]$$

4. Schritt

Wir zeigen nun, dass die beiden letzen Summen jeweils den Wert 0 haben. Zunächst der zweit letzte Term:

$$\begin{aligned} & \sum_{i=1}^{n} 2 \cdot k \cdot [y_i - \overline{y}] \\ & = 2 \cdot k \cdot [y_1 - \overline{y}] + ... + 2 \cdot k \cdot [y_n - \overline{y}] \\ & = 2 \cdot k \cdot y_1 - 2 \cdot k \cdot \overline{y} + ... + 2 \cdot k \cdot y_n - 2 \cdot k \cdot \overline{y} \end{aligned}$$

Der zuletzt notierte Term wird wieder umgeordnet, indem man die Summanden, die die Messwerte y_i enthalten, nebeneinander schreibt und die restlichen n (!) Summanden der Form $2 \cdot k \cdot \overline{y}$ zusammen zählt, was $n \cdot 2 \cdot k \cdot \overline{y}$ ergibt. Man erhält

$$2 \cdot k \cdot (y_1 + ... + y_n) - n \cdot 2 \cdot k \cdot \overline{y} .$$

Wir setzen nun für den Mittelwert $\overline{y}$ der Messwerte y_i den entsprechenden Term

$$\overline{y} = \frac{1}{n} \cdot (y_1 + ... + y_n)$$

ein und sehen (Kürzen durch n), dass der Term tatsächlich den Wert 0 hat:

$$2 \cdot k \cdot (y_1 + ... + y_n) - n \cdot 2 \cdot k \cdot \frac{1}{n} \cdot (y_1 + ... + y_n) = 0.$$

Dass der letzte Term den Wert 0 hat, zeigt man analog.

Es ist auch ohne viel Umformungen einsichtig, dass diese beiden Summen den Wert 0 haben müssen: Man summiert die jeweiligen Abweichungen (mit Vorzeichen!) aller Messwerte vom Mittelwert auf, ohne jeweils die Beträge oder Quadrate zu nehmen, was 0 ergeben muss.

Somit erhält man für Q(a; b)

$$Q(a;b) = \sum_{i=1}^{n} [y_i - \overline{y}]^2 + \sum_{i=1}^{n} a^2 \cdot [x_i - \overline{x}]^2 + \sum_{i=1}^{n} (-2 \cdot a)[y_i - \overline{y}] \cdot [x_i - \overline{x}] \quad + n \cdot k^2$$

5. Schritt

Es werden nun einige Abkürzungen eingeführt, denn die drei ersten Summen sind „alte Bekannte“:

- Die erste Summe ist das n-fache der Varianz V_y der y-Messwerte:

$$n \cdot V_y = n \cdot \frac{1}{n} \cdot \left([y_1 - \overline{y}]^2 + ... + [y_n - \overline{y}]^2 \right)$$
$$= \left([y_1 - \overline{y}]^2 + ... + [y_n - \overline{y}]^2 \right).$$

- Die zweite Summe ist a^2 mal das n-fache der Varianz V_x der x-Messwerte und die dritte Summe ist (-2a)-mal das n-fache der Kovarianz c_{xy}.

Somit hat man (knapp notiert):

$$Q(a;b) = n \cdot V_y + a^2 \cdot n \cdot V_x - 2 \cdot a \cdot n \cdot c_{xy} + n \cdot k^2$$

Erst einmal schön. Ruhe.

Wir erinnern uns daran, wohin wir eigentlich steuern möchten, und machen uns die Sachlage bewusst:

- Das Ziel ist, a und b so zu bestimmen, dass Q(a; b) minimal ist.
- Der Ordinatenabschnitt b - und auch a - sind im letzen Summand in k ($k = \overline{y} - a \cdot \overline{x} - b$) enthalten, während die mittleren zwei Summanden nur a enthalten.
- Alle restlichen Symbole (V_y, V_x, c_{xy}, n, $\overline{y}$ und $\overline{x}$)stehen für Zahlen, deren Werte man nach der statistischen Erhebung schon berechnet hat und somit kennt.

Jetzt wird zugeschlagen!

6. Schritt

Wir betrachten zunächst die mittleren zwei Summen von Q(a; b), und bestimmen zunächst a so, dass der Wert dieses Terms minimal ist.

$$a^2 \cdot n \cdot V_x - 2 \cdot a \cdot n \cdot c_{xy}.$$

Um zu erkennen, wie a gewählt werden muss, damit dieser Term minimal ist, formen wir diese Summe mit Hilfe einer quadratischen Ergänzung um:

$$
\begin{aligned}
& a^2 \cdot n \cdot V_x - 2 \cdot a \cdot n \cdot c_{xy} \\
&= n \cdot V_x \cdot (a^2 - 2 \cdot a \cdot \frac{c_{xy}}{V_x}) \\
&= n \cdot V_x \cdot \left(\left[a - \frac{c_{xy}}{V_x} \right]^2 - \left(\frac{c_{xy}}{V_x} \right)^2 \right) \\
&= n \cdot V_x \cdot \left[a - \frac{c_{xy}}{V_x} \right]^2 - n \cdot \frac{c_{xy}^{\ 2}}{V_x}
\end{aligned}
$$

Dieser Term ist dann minimal, wenn gilt:

$$\boxed{a = \frac{c_{xy}}{V_x}.}$$

Denn dann ist der Wert der quadrierten Klammer 0, und kleiner geht es nicht, da die Werte des Quadrates ja immer größer oder eben gleich 0 sind.

Unsere Summe Q(a; b) hat sich somit reduziert auf den Term

$$Q(a;b) = n \cdot V_y - n \cdot \frac{c_{xy}^2}{V_x} + n \cdot k^2$$

Nun wählen wir b - in Abhängigkeit von a - so, dass k=0 ist.
Betrachte k.

$$k = \overline{y} - a \cdot \overline{x} - b$$

Wenn der Ordinatenabschnitt b der Geraden so gewählt wird, dass

$$\boxed{\begin{aligned} b &= \overline{y} - a \cdot \overline{x} \\ &= \overline{y} - \frac{c_{xy}}{V_x} \cdot \overline{x} \end{aligned}}$$

dann ist k=0, und der letzte Summand von Q(a; b) ist minimal, da ja $n \cdot k^2$ nicht kleiner als 0 sein kann.

Die Summe Q(a; b) reduziert sich schlussendlich auf

$$Q(a;b) = n \cdot V_y - n \cdot \frac{c_{xy}^2}{V_x} = n \cdot V_y \cdot \left(1 - \frac{c_{xy}^2}{V_x \cdot V_y} \right),$$

oder mit Hilfe des Korrelationskoeffizienten $r = \frac{c_{xy}}{s_x \cdot s_y}$:

$$\boxed{Q(a;b) = n \cdot V_y \cdot (1 - r^2).}$$

Name:_____ Datum:_____

Das Ziel ist erreicht.
Steigungsfaktor a und Ordinatenabschnitt b des Terms der Regressionsgeraden sind bestimmt, für die Summe der quadratischen Abweichungen Q(a;b) der Regressionsgeraden haben wir einen übersichtlichen Ausdruck hergeleitetet.

Mit der Standardabweichung $s = \sqrt{V}$ und dem Korrelationskoeffizienten r ergibt sich nach einigen einfachen Umformungen für den Term der Regressionsgerade

$$\begin{aligned} y(x) &= a \cdot x + b \\ &= \underbrace{\frac{c_{xy}}{V_x}}_{a} \cdot x + \underbrace{\bar{y} - \frac{c_{xy}}{V_x} \cdot \bar{x}}_{b} \\ &= \underbrace{r \cdot \frac{s_y}{s_x}}_{a} \cdot x + \underbrace{\bar{y} - r \cdot \frac{s_y}{s_x} \cdot \bar{x}}_{b}. \end{aligned}$$

Die Summe der Quadratischen Abweichungen bei der Regressionsgeraden ist

$$Q(a;b) = n \cdot V_y \cdot (1 - r^2).$$

Zur Vorgehensweise bei der Bestimmung der relevanten Werte der Regressionsgeraden:

- Berechne nacheinander $\bar{x}, \bar{y}, V_x, V_y, s_x, s_y, c_{xy}$ und schließlich r.
- Berechne mit Hilfe der obigen Formeln die Steigung a und den Ordinatenabschnitt b sowie die Summe der quadratischen Abweichungen.