Name: **Datum:**

Herleitung des Terms der Regressionsgeraden II

Wir gehen aus von einer bivariaten statistischen Erhebung mit

- einer Grundgesamtheit mit dem Erhebungsumfang n,
- zwei quantitativen Merkmalen X und Y,
- der durch die Erhebung gewonnen Urliste mit den Messwertepaaren $(x_1 \mid y_1), (x_2 \mid y_2), ..., (x_n \mid y_n)$,
- den Arithmetischen Mittelwerten $\bar{x}$ und $\bar{y}$,
- den Mittleren Quadratische Abweichungen oder Varianzen V_X und V_Y,
- den Standardabweichungen s_X und s_Y,
- der Kovarianz c_{xy} sowie
- dem Korrelationskoeffizienten r.

Gesucht sind die Werte des **Steigungsfaktors a** und des **Ordinatenabschnitts b** des Terms $y(x) = a \cdot x + b$ der Regressionsgeraden. Dabei ist der Begriff Regressionsgerade über folgende Bedingung festgelegt:

Der **Steigungsfaktor a** und der **Ordinatenabschnitt b** des **Terms der Regressionsgeraden** sollen so bestimmt werden, dass die **Summe der Abweichungsquadrate** zwischen den **Messwerten y_i** und den entsprechenden **Funktionswerten $y(x_i)$ der Regressionsgeraden** $y(x) = a \cdot x + b$, also

$$\begin{aligned} Q(a;b) &= (y(x_1) - y_1)^2 + ... + (y(x_n) - y_n)^2 \\ &= ((a \cdot x_1 + b) - y_1)^2 + ... + ((a \cdot x_n + b) - y_n)^2 \end{aligned}$$

minimal wird.

Beachte: Der obige Term $Q(a;b)$ enthält lediglich 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, und wie groß dann diese Summe der Abweichungsquadrate für die Regressionsgerade ist.

1.Schritt:

Kompliziertere Rechnungen, die wir an dieser Stelle nicht durchführen wollen, ergeben, dass der sogenannte Schwerpunkt $S(\bar{x} \mid \bar{y})$ auf der Regressionsgeraden liegt, d.h. es gilt

$$y(\bar{x}) = \bar{y} \Leftrightarrow a \cdot \bar{x} + b = \bar{y} \Leftrightarrow b = \bar{y} - a \cdot \bar{x}$$

Setzt man nun den so erhaltenen Term für b in den Term $y(x) = a \cdot x + b$ der Regressionsgerade ein und vereinfacht den Term, so ergibt sich

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

Man sieht, dass somit nur noch der **Steigungsfaktor a** des Terms der Regressionsgerade bestimmt werden muss und sich der Term $Q(a;b)$ vereinfacht zu

$$Q(a) = (a \cdot (x_1 - \bar{x}) + \bar{y} - y_1)^2 + ... + (a \cdot (x_n - \bar{x}) + \bar{y} - y_n)^2$$

2.Schritt:

Wir ordnen nun die einzelnen Summanden in den Klammern etwas um, fassen anders zusammen und quadrieren die Klammern nach der 2.Binomischen Formel aus:

$$\begin{aligned} Q(a) &= (a\cdot(x_1-\bar{x})+\bar{y}-y_1)^2+\ldots+(a\cdot(x_n-\bar{x})+\bar{y}-y_n)^2 \\ &= (a\cdot(x_1-\bar{x})-(y_1-\bar{y}))^2+\ldots+(a\cdot(x_n-\bar{x})-(y_n-\bar{y}))^2 \\ &= [a^2\cdot(x_1-\bar{x})^2-2\cdot a\cdot(x_1-\bar{x})\cdot(y_1-\bar{y})+(y_1-\bar{y})^2]+\ldots \\ &\qquad \ldots+[a^2\cdot(x_n-\bar{x})^2-2\cdot a\cdot(x_n-\bar{x})\cdot(y_n-\bar{y})+(y_n-\bar{y})^2] \end{aligned}$$

Nun sortieren wir die Summanden um und klammern a^2 bzw. $-2a$ aus:

$$\begin{aligned} Q(a) &= a^2(x_1-\bar{x})^2+\ldots+a^2(x_n-\bar{x})^2 \\ &\qquad -2a(x_1-\bar{x})(y_1-\bar{y})-\ldots-2a(x_n-\bar{x})(y_n-\bar{y}) \\ &\qquad +(y_1-\bar{y})^2+\ldots+(y_n-\bar{y})^2 \\ &= a^2[(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2] \\ &\qquad -2a[(x_1-\bar{x})(y_1-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})] \\ &\qquad +[(y_1-\bar{y})^2+\ldots+(y_n-\bar{y})^2] \end{aligned}$$

Nun sind

- $[(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2]=n\cdot V_X$,
- $[(x_1-\bar{x})(y_1-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})]=n\cdot c_{XY}$ und
- $[(y_1-\bar{y})^2+\ldots+(y_n-\bar{y})^2]=n\cdot V_Y$,

so dass der Term $Q(a)$ mit diesen Bezeichnungen zuerst einmal wesentlich einfacher aussieht.

$$Q(a) = a^2\cdot n\cdot V_X-2a\cdot n\cdot c_{XY}+n\cdot V_Y$$

Beachte wieder: Der obige Term $Q(a)$ enthält lediglich die Variable a; die übrigen „Formvariablen“ n, V_X, c_{xy} und V_Y muss man sich mit den entsprechenden Werten belegt vorstellen.

Der Term $Q(a)$ ist – wie man jetzt deutlich erkennen kann – ein Quadratischer Term in der Variablen a. Da wir den Wert für den Steigungsfaktor a suchen, bei dem der Term $Q(a)$ <u>minimal</u> wird, formen wir den Term $Q(a)$ mittels Quadratischer Ergänzung in die <u>Scheitelpunktform</u> um:

$$\begin{aligned} Q(a) &= n\cdot V_X\cdot\left[a^2-2a\cdot\frac{c_{XY}}{V_X}+\frac{V_Y}{V_X}\right] \\ &= n\cdot V_X\cdot\left[a^2-2a\cdot\frac{c_{XY}}{V_X}+\left(\frac{c_{XY}}{V_X}\right)^2-\left(\frac{c_{XY}}{V_X}\right)^2+\frac{V_Y}{V_X}\right] \\ &= n\cdot V_X\cdot\left[\left(a-\frac{c_{XY}}{V_X}\right)^2-\frac{{c_{XY}}^2}{{V_X}^2}+\frac{V_Y}{V_X}\right] \\ &= n\cdot V_X\cdot\left(a-\frac{c_{XY}}{V_X}\right)^2-\frac{n\cdot{c_{XY}}^2}{V_X}+n\cdot V_Y \end{aligned}$$

Ergebnis:

Jetzt sieht man, dass der Term $Q(a)$ seinen minimalen Wert für $\boxed{a = \frac{c_{XY}}{V_X}}$ annimmt und wir somit den gesuchten Steigungsfaktor a der Regressionsgerade gefunden haben.

Erinnern wir uns daran, dass sich der Ordinatenabschnitt b aus dem Steigungsfaktor a durch $b = \bar{y} - a \cdot \bar{x}$ errechnet, so haben wir auch den Ordinatenabschnitt $\boxed{b = \bar{y} - \frac{c_{XY}}{V_X} \cdot \bar{x}}$.

Der Term der Regressionsgerade lautet dann $\boxed{y(x) = \underbrace{\frac{c_{XY}}{V_X}}_{a} \cdot x + \underbrace{\bar{y} - \frac{c_{XY}}{V_X} \cdot \bar{x}}_{b}}$.

Der Term $Q(a;b)$, d.h. die Summe der Abweichungsquadrate reduziert sich schlussendlich auf

$$Q(a;b) = n \cdot V_Y - \frac{n \cdot c_{XY}^{\,2}}{V_X} = n \cdot V_Y \cdot \left(1 - \frac{c_{XY}^{\,2}}{V_X \cdot V_Y}\right),$$

oder mit Hilfe des Korrelationskoeffizienten $r = \frac{c_{XY}}{s_X \cdot s_Y}$ auf $\boxed{Q(a;b) = n \cdot V_Y \cdot (1 - r^2).}$

Damit haben wir unser Ziel 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.

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.