Beschreibende Statistik Mittelwert

Unter dem arithmetischen Mittel (Mittelwert) $\overline{x}$ von n Zahlen verstehen wir:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

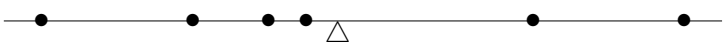
Diesen Mittelwert untersuchen wir etwas genauer.

1. Zeige für $n = 3$:

   $$\sum_{i=1}^{n} (\overline{x} - x_i) = 0$$

   d. h. die Summe der Abweichungen vom Mittelwert ist Null.

2. $\overline{x}$ kann als Schwerpunkt interpretiert werden, erläutere dies.



$$\sum_{x_i<x_s} (x_s - x_i) \;=\; \sum_{x_s<x_i} (x_i - x_s)$$

$$\Longrightarrow \quad x_s = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3. $x = \overline{x}$ minimiert die quadratische Funktion.

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

   d. h. die Summe der quadratischen Abweichungen wird für $x = \overline{x}$ minimal.
   Diese Eigenschaft benötigen wir, um die Regressionsgerade zu ermitteln.

   Da eine Parabel vorliegt, wird der x-Wert des Scheitels berechnet, und zwar durch Bestimmung von Nullstellen, wobei der konstante Summand (Verschiebung in y-Richtung) entfallen kann.

   $$\begin{array}{lll} y = ax^2 + bx + c & & \\ y = ax^2 + bx & 0 = ax^2 + bx & \\ & 0 = x(ax + b) & \\ & x_1 = 0 \qquad x_2 = -\frac{b}{a} & \Longrightarrow \quad x_{\text{Scheitel}} = -\frac{b}{2a} \end{array}$$

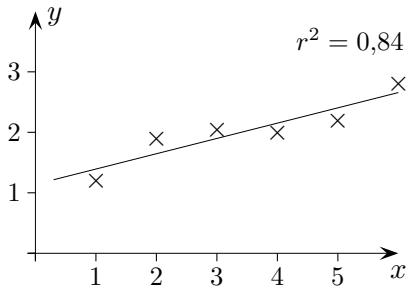
   Nun ist

   $$f(x) = \sum_{i=1}^{n} (x - x_i)^2 = \sum_{i=1}^{n} (x^2 - 2xx_i + x_i^2) = nx^2 - 2x\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2$$

   $$\Longrightarrow \quad x_{\text{Scheitel}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Regressionsgerade

x	x_1	x_2	x_3	$\dots$	x_n
y	y_1	y_2	y_3	$\dots$	y_n



Beim Auswerten von Messreihen wird häufig eine durch theoretische Überlegungen nahegelegte lineare Beziehung zwischen den x- und y-Werten gesucht, d. h. eine Gerade $y = mx + b$ (Regressionsgerade), die die Datenpunkte möglichst gut approximiert.

Als Abweichungsmaß kann (nach Gauss) die Summe der Quadrate der Differenzen

$$Q = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + \dots + (mx_n + b - y_n)^2$$

genommen werden. Hierbei werden m und b so gewählt, dass Q einen kleinsten Wert annimmt.

Q kann als quadratische Funktion der Variablen m bzw. b betrachtet werden.

$$Q = \sum (mx_i + b - y_i)^2$$ Die Summe erstreckt sich stets von 1 bis n.

$$Q(b) = \sum (b - (y_i - mx_i))^2$$ siehe 3., vorherige Seite

$$\Longrightarrow \quad b_{\min} = \frac{1}{n} \sum_{i=1}^{n} (y_i - mx_i)$$

$$\Longrightarrow \quad \overline{y} = m\overline{x} + b$$ Mittelwerte $\overline{x} = \frac{1}{n}\sum x_i$, $\overline{y} = \frac{1}{n}\sum y_i$
$P(\overline{x} \mid \overline{y})$ liegt auf der Ausgleichsgeraden.

m ist noch zu bestimmen.

Um die Rechnung einfach zu halten, wählen wir den Schwerpunkt als Ursprung:

$d_i = x_i - \overline{x}$
$e_i = y_i - \overline{y}$

b ist dann Null, die Steigung hat sich nicht verändert.

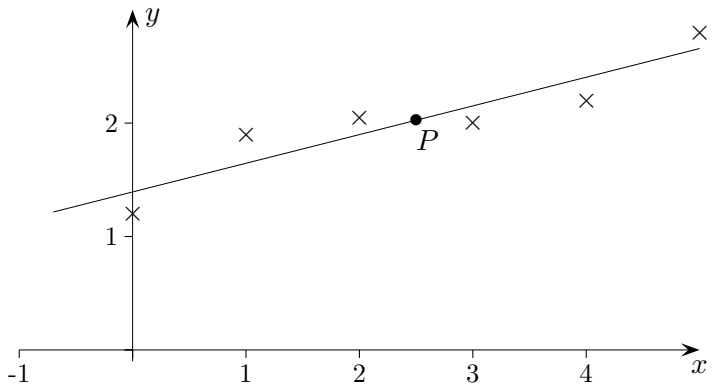
$$\begin{aligned} Q(m) &= \sum (md_i - e_i)^2 \\ &= \sum (m^2 d_i^2 - 2md_i e_i + e_i^2) \\ &= m^2 \sum d_i^2 - 2m \sum d_i e_i + \sum e_i^2 \end{aligned}$$

$$\Longrightarrow \quad m_{\min} = \frac{\sum d_i e_i}{\sum d_i^2}$$ siehe 3., vorherige Seite

$$m = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \qquad \left(= \frac{\frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2} = \frac{Cov_{xy}}{\sigma_x^2} \right)$$

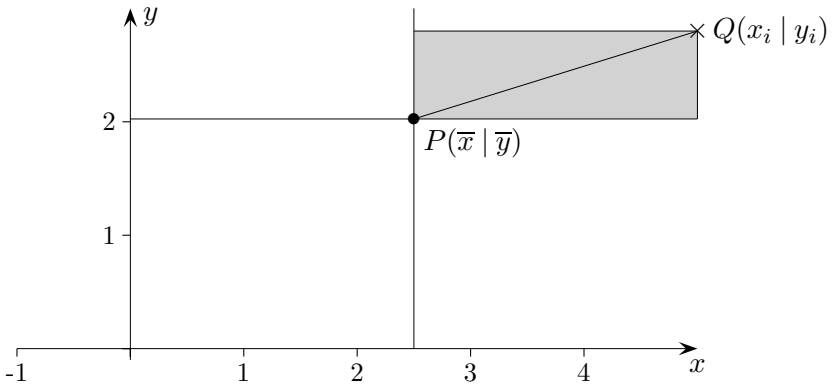
Die Gleichung der Regressionsgeraden lautet daher: $y = m(x - \overline{x}) + \overline{y}$

Regressionsgerade $y = m(x - \overline{x}) + \overline{y}$, anschaulich



Es erscheint plausibel, dass der Schwerpunkt $P(\overline{x} \mid \overline{y})$ auf der Ausgleichsgeraden liegt,
$\overline{x} = \frac{1}{n}\sum x_i$, $\overline{y} = \frac{1}{n}\sum y_i$.

Aber auch die Steigung $m = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$
kann veranschaulicht werden.



Hierzu betrachte man den Term $\dfrac{(x_i - \overline{x})(y_i - \overline{y})}{(x_i - \overline{x})^2}$, den Inhalt des grauen Rechtecks, $\dfrac{\Delta y}{\Delta x}$
und fasse die Summenbildung als eine Art Durchschnittsbildung auf.

Korrelationskoeffizient

Wir benötigen ein Maß dafür, wie stark die Datenpunkte um die Regressionsgerade streuen.
Dazu rechnen wir die quadratische Abweichung aus.

$$
\begin{aligned}
Q &= \sum (e_i - m d_i)^2 && m = \frac{\sum e_i d_i}{\sum d_i^2} \\
&= \sum (e_i^2 - 2 e_i m d_i + (m d_i)^2) \\
&= \sum e_i^2 - 2m \sum d_i e_i + m^2 \sum d_i^2 \\
&= \sum e_i^2 - \frac{2(\sum e_i d_i)^2}{\sum d_i^2} + \frac{(\sum e_i d_i)^2 \sum d_i^2}{(\sum d_i^2)^2} && \text{kürzen durch } \sum d_i^2 \\
&= \sum e_i^2 - \frac{(\sum e_i d_i)^2}{\sum d_i^2}
\end{aligned}
$$

Je kleiner der Term $\frac{(\sum e_i d_i)^2}{\sum d_i^2}$ ist, desto größer ist die Quadratsumme.

Diese ist Null, falls $\frac{(\sum e_i d_i)^2}{\sum d_i^2} = \sum e_i^2$ ist.

Da $Q \geq 0$ ist, folgt $\quad 0 \leq \frac{(\sum e_i d_i)^2}{\sum d_i^2} \leq \sum e_i^2$

$$\Longrightarrow \qquad 0 \leq \frac{(\sum e_i d_i)^2}{\sum d_i^2 \sum e_i^2} \leq 1$$

Der mittlere Term heißt *Bestimmtheitsmaß*.

Gebräuchlicher ist der *Korrelationskoeffizient*

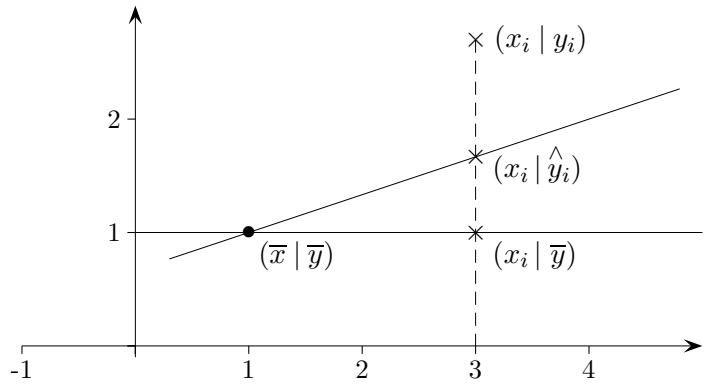
$$r = \frac{\sum e_i d_i}{\sqrt{\sum d_i^2 \sum e_i^2}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \qquad \left(= \frac{\frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \right)$$

r ist die Wurzel aus dem Bestimmtheitsmaß
und hat im Gegensatz zu diesem stets dasselbe Vorzeichen wie die Steigung m (leicht zu sehen).
Es ist: $\quad -1 \leq r \leq 1$.

Für $r = 0$ ist die Steigung m auch Null. Es liegt kein linearer Zusammenhang vor,
für $r = 1$ und $r = -1$ liegen die Datenpunkte auf der Regressionsgeraden.
Mit den Termen für m und r erhalten wir unmittelbar den Zusammenhang: $m = r \cdot \frac{\sigma_y}{\sigma_x}$.
Zu beachten ist, dass ein hoher Korrelationskoeffizient nicht eine *kausale* Abhängigkeit bedeuten muss.

In Excel können Ausgleichsgeraden ohne Aufwand ausgegeben werden:
Auf einen Datenpunkt klicken, mit rechter Maustaste Trendlinie hinzufügen, Trendlinie formatieren,
Optionen, Gleichung und Bestimmtheitsmaß im Diagramm darstellen.

Bestimmtheitsmaß



Um die Güte der Regression zu beurteilen, sind die Differenzen $y_i - \hat{y}_i$ zu betrachten,
mit $\hat{y}_i = m(x_i - \overline{x}) + \overline{y}$.

Nun gilt:

$$* \quad (y_i - \overline{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

Je besser die Regression, umso kleiner ist $(y_i - \hat{y}_i)$ oder
um so mehr gleicht sich $(\hat{y}_i - \overline{y})$ der Abweichung $(y_i - \overline{y})$ an.

Die Beziehung $*$ kann quadriert und aufsummiert werden. Dies ergibt:

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

Hierbei ist zu beachten, dass der entstehende mittlere Term $2 \cdot \sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$ null ergibt.
Durch Umformen und Einsetzen der Steigung der Regressionsgeraden kann dies überprüft werden.

Das Bestimmtheitsmaß B ist nun der Anteil von $\sum (\hat{y}_i - \overline{y})^2$ an der Gesamtstreuung $\sum (y_i - \overline{y})^2$.

$$B = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$$ (= Streuung, die sich aus der Regression ergibt / Gesamtstreuung)

oder

$$B = \frac{\sum (y_i - \overline{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}$$ (= 1− Streuung, die trotz Regression verbleibt / Gesamtstreuung)

Es kann nun gezeigt werden, dass B mit r^2 übereinstimmt.
Aus der Rechnung auf der vorigen Seite entnehmen wir:

$$\sum (e_i - m d_i)^2 = \sum e_i^2 - \frac{(\sum e_i d_i)^2}{\sum d_i^2} \qquad | : \sum e_i^2$$

$$\frac{\sum (e_i - m d_i)^2}{\sum e_i^2} = 1 - \frac{(\sum e_i d_i)^2}{\sum e_i^2 \sum d_i^2}$$

$$\underbrace{\frac{(\sum e_i d_i)^2}{\sum e_i^2 \sum d_i^2}}_{r^2} = \underbrace{1 - \frac{\sum (e_i - m d_i)^2}{\sum e_i^2}}_{B}$$

$d_i = x_i - \overline{x}$
$e_i = y_i - \overline{y}$

In der Korrelationsanalyse wird der zumeist lineare Zusammenhang zweier Merkmale untersucht. Ein Maß für die Stärke der Abhängigkeit ist der Korrelationskoeffizient.

In der Regressionsanalyse wird die Art des Zusammenhangs durch eine Funktion erfasst.