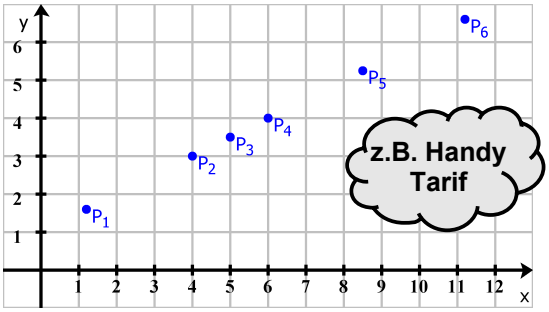
Regressionsgerade

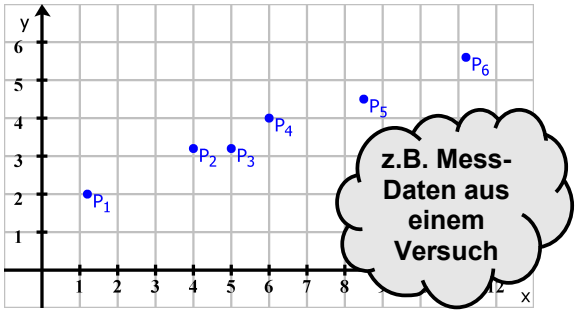
Wo ist das Problem?

Beispiel 1:



Das Schaubild zeigt eine Punktewolke von 6 Punkten. Es ist offensichtlich, dass alle auf der Geraden mit der Gleichung $y = 0{,}5\,x + 1{,}5$ liegen.

Beispiel 2:



Die Punkte in dem zweiten Schaubild liegen dagegen nicht auf einer Geraden.

Für das zweite Beispiel ist eine Gerade gesucht, für die die Summe der Abstände zu den einzelnen Punkten möglichst klein ist.

Eine solche Gerade ist z.B. eine **Regressionsgerade**.

Der magische Punkt Q

Auf jeder Regressionsgeraden g liegt ein Punkt Q, dessen Koordinaten sich aus den Mittelwerten der Koordinaten der Punkte berechnet.

Beispiel: $P_1(2\,|\,1);\ P_2(5\,|\,4);\ P_3(11\,|\,4) \ \Rightarrow\ Q\left(\frac{2+5+11}{3}\,\middle|\,\frac{1+4+4}{3}\right) \ \Rightarrow\ Q(6\,|\,3)$

Q ist die halbe Miete

Für Regressionsgerade gilt: $y = m(x - x_Q) + y_Q$, wobei x_Q und y_Q die Koordinate von Q sind.

Beispiel: $Q(6\,|\,3) \ \Rightarrow\ y = m(x-6)+3$

Das Problem mit dem Abstand...

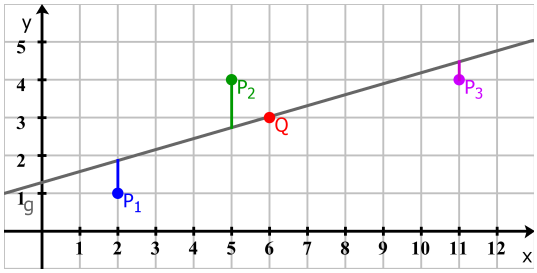
Die Abstände zur Regressionsgeraden sollen in der Summe möglichst klein sein:

Der Abstand d zwischen Punkt und Gerade berechnet sich mit

$d = m(x_P - x_Q) + y_Q - y_P$, wenn der Punkt unterhalb der Geraden liegt

$d = y_P - \left(m(x_P - x_Q) + y_Q\right)$, wenn der Punkt oberhalb der Geraden liegt

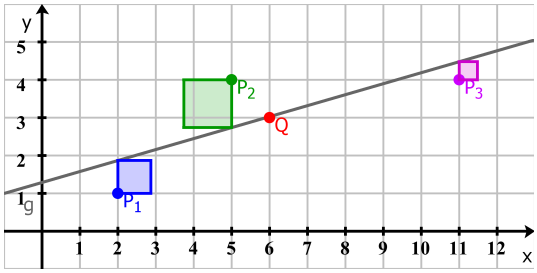
(Abstände sind immer positiv)



Wir wissen aber doch gar nicht, ob ein Punkt oberhalb oder unterhalb der Regressionsgeraden liegt!

… und dessen Lösung

Statt der Abstände können wir genauso gut deren Quadrate betrachten:



Werden die Abstände größer, so werden auch die Quadrate größer und umgekehrt. Die Quadrate haben folgenden Vorteil: Deren Flächeninhalt berechnet sich wie folgt

$$A_{\text{Quadrat}}=\left(m(x_P-x_Q)+y_Q-y_P\right)^2=\left(y_P-\left(m(x_P-x_Q)+y_Q\right)\right)^2$$

es spielt keine Rolle mehr, ob der Punkt ober- oder unterhalb der Geraden liegt.

Beispiel:

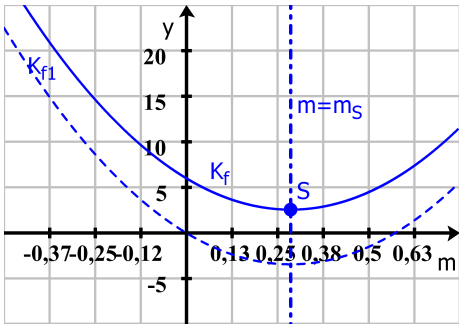
$Q(6\,|\,3)$

$$\begin{array}{lllll}
P_1(2\,|\,1) & \Rightarrow & y=\left(m(2-6)+3-1\right)^2 & = & y=\left(1-\left(m(2-6)+3\right)\right)^2 \\
P_2(5\,|\,4) & \Rightarrow & y=\left(m(5-6)+3-4\right)^2 & = & y=\left(4-\left(m(5-6)+3\right)\right)^2 \\
P_3(11\,|\,4) & \Rightarrow & y=\left(m(11-6)+3-4\right)^2 & = & y=\left(4-\left(m(11-6)+3\right)\right)^2
\end{array}$$

$$\begin{array}{lll}
 & = & 16m^2-16m+4 \\
 & = & m^2+2m+1 \\
 & = & 25m^2-10m+1 \\
+ & & \\
\text{(Summe der Quadrate)} & = & 42m^2-24m+6
\end{array}$$

Woher m nehmen, wenn nicht stehlen?

Die Summe der Quadrate soll möglichst klein sein ⇒ es wird ein m gesucht, dass $42m^2-24m+6$ möglichst klein ist.



K_f ist der Graph von $f(m)=42m^2-24m+6$. K_f ist eine nach oben geöffnete Parabel. $f(m)$ Nimmt ihren kleinsten Wert im Scheitelpunkt an.

Berechnung des Scheitelpunktes:

Bilde $f_1(m)=42m^2-24m$. Die x-Koordinaten der Scheitelpunkte von K_{f_1} und K_f sind gleich.

Bestimme die Nullstellen von $f_1(m)$: Setze $f_1(m)=0 \Rightarrow m=0 \vee m=\frac{4}{7}$

Die x-Koordinate des Scheitelpunktes ist bei $m_S=\frac{\frac{4}{7}}{2}=\frac{2}{7}$

Die gesuchte Steigung ist somit $m=\frac{2}{7}$

Die Regressionsgerade

$g: y=\frac{2}{7}(x-6)+3=\frac{2}{7}x+\frac{9}{7}$ ist die Gleichung der gesuchten Regressionsgeraden.