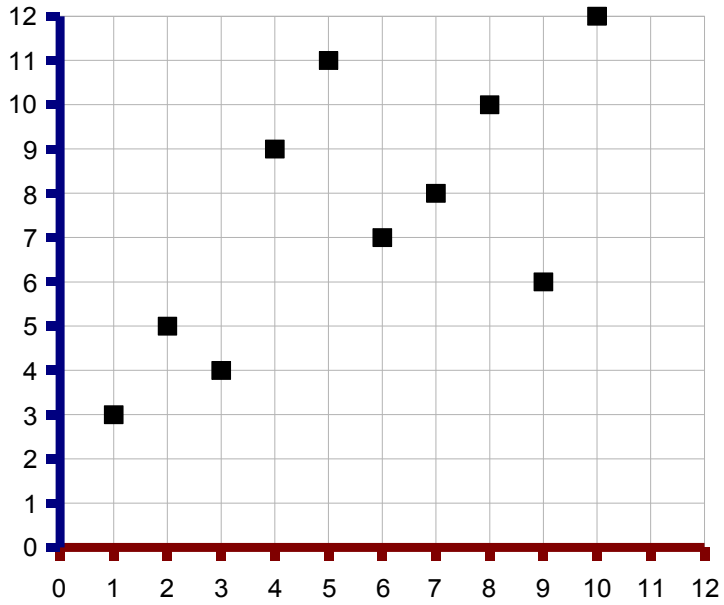
Streudiagramme

für 2 metrisch skalierte Variablen

paarweise Messwerte (x,y)

x	3	10	6	4	8	2	5	7	9	1
y	4	12	7	9	10	5	11	8	6	3



Aussagen zu Zusammenhängen

1.

empirische Kovarianz *Standardabweichung der WertPAARE*	s_{xy}

$$s_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\frac{(Wert\ x_1-Mittelwert\ aller\ x)(Wert\ y_1-Mittelwert\ aller\ y)+(Wert\ x_2-Mittelwert\ aller\ x)(Wert\ y_2-Mittelwert\ aller\ y)+...}{Anzahl\ der\ WertPAARE}$$

x	3	10	6	4	8	2	5	7	9	1
y	4	12	7	9	10	5	11	8	6	3

$\bar{x}=5{,}5$

$\bar{y}=7{,}5$

$$s_{xy}=\frac{1}{10}(3-5{,}5)(4-7{,}5)+(10-3{,}5)(12-7{,}5)+(4-3{,}5)(9-7{,}5)+....\approx 5{,}55$$

Erinnerung: Standardabweichung

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

2.

empirischer Korrelationskoeffizient Bravais-Pearson

"normierte" Standardabweichung der WertPAARE

$$r = r_{xy}$$

Der Teil 1/n aus den „Einzelformeln" kürzt sich weg

Normierung

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\frac{\text{Abweichung zusammen}}{\text{Abweichung x} \cdot \text{Abweichung y}}$$

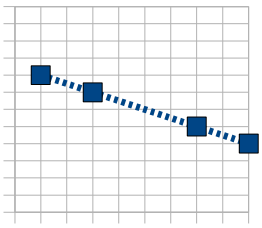
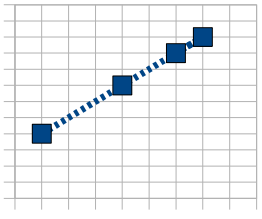
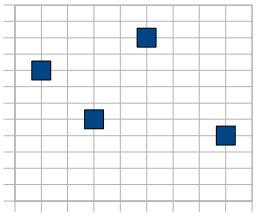
x	3	10	6	4	8	2	5	7	9	1
y	4	12	7	9	10	5	11	8	6	3

$s_x \approx 2{,}87$ $s_y \approx 2{,}87$ $s_{xy} \approx 5{,}55$

$$r_{xy} = \frac{5{,}55}{2{,}87 \cdot 2{,}87} \approx 0{,}67$$

r kann **Werte zwischen -1 und +1** annehmen,
und gibt damit Auskunft über **Stärke und Richtung des Zusammenhangs**

Interpretation:

r = -1	vollständiger negativer Zusammenhang (je mehr, desto weniger) *alle Punkte liegen auf einer fallenden Geraden*	
r = 1	vollständiger positiver Zusammenhang (je mehr, desto mehr) *alle Punkte liegen auf einer steigenden Geraden*	
r = 0	kein **LINEARER** Zusammenhang *...vielleicht aber ein nicht-linearer (exponentieller, u-förmiger oder ...)*	

keine Aussage zu kausalen Beziehungen!

Scheinkorrelationen möglich!

je höher der Absatz von Winterreifen, desto höher der Absatz von Nikoläusen...

3.

lineare Regression

Regressionsgerade

...auf der Suche nach
einem berechenbaren Zusammenhang von x und y
(in Form einer linearen Gleichung /Funktion)

mit dem Ziel,
anhand der Ausprägung eines (beliebigen) x-Wertes
die Ausprägung des entsprechenden y-Wertes
(ungefähr) berechnen zu können

Beziehung zwischen
Unabhängiger **V**ariable + **A**bhängiger **V**ariable?

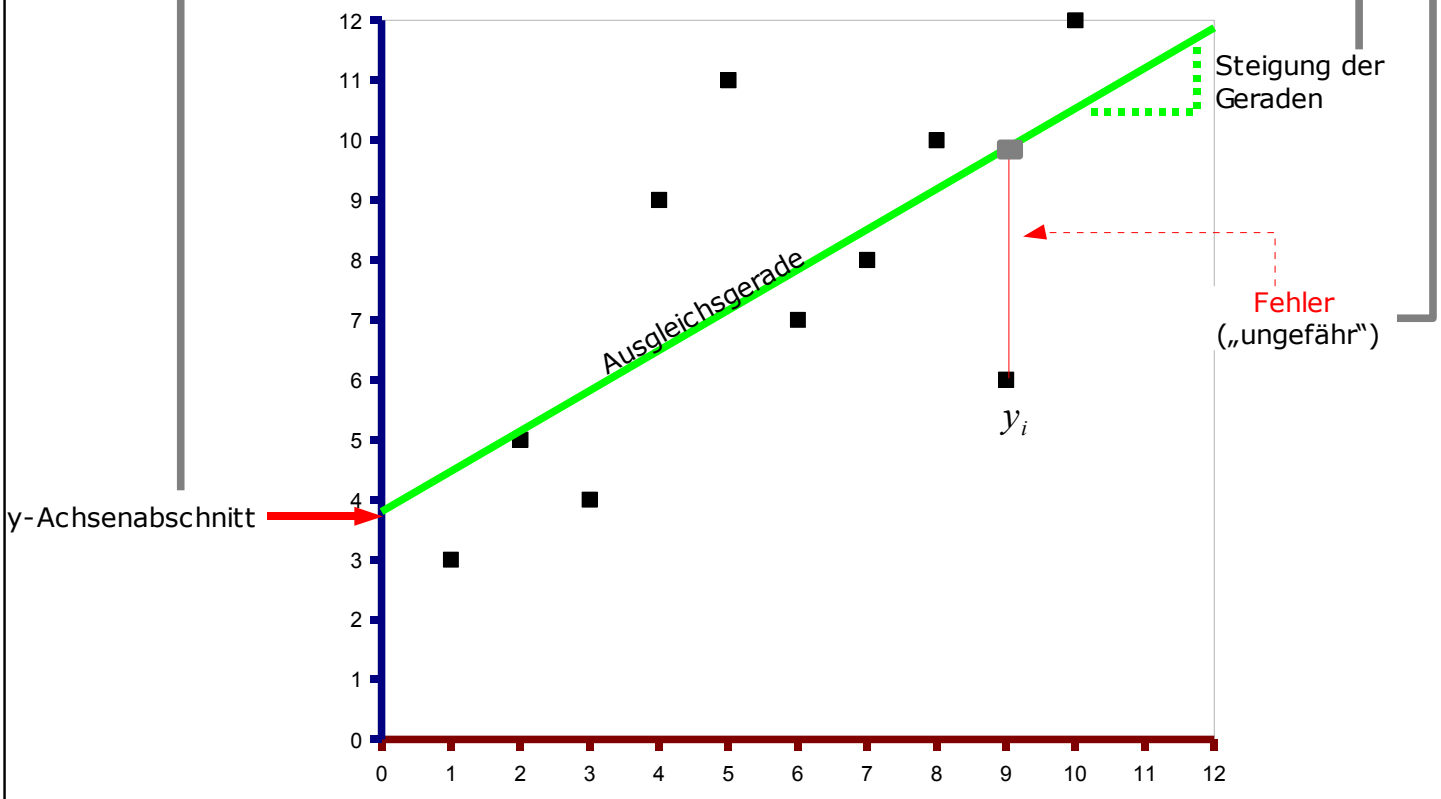
Gleichung

$$y_i = \alpha + \beta \cdot x_i + e_i$$

Funktion

$$f(x) = \alpha + \beta \cdot x$$

$$y_i = \alpha + \beta \cdot x_i + e_i$$



Berechnung der Ausgleichs-/Regressionsgeraden

$$f(x) = \alpha + \beta \cdot x$$

Kleinste-Quadrate-Schätzer

...auf der Suche nach
einem Achsenabschnitt a und einer Steigung β

mit dem Ziel,
den Gesamtfehler (e1+e2+e3+...) so klein wie möglich zu halten

-> Minimierung
*der **durchschnittlichen Abweichung** der tatsächlichen Werte*
von den Werten auf der Ausgleichsgeraden

$\hat{\alpha}$ $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

vergleiche:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

wegen irgendwelchen Wegkürzungen...

$s_{xx} \mathrel{\widehat{=}} s_x^2$

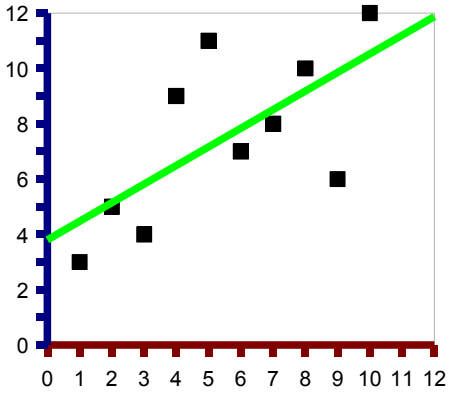
x	3	10	6	4	8	2	5	7	9	1	$\bar{x}=5{,}5$
y	4	12	7	9	10	5	11	8	6	3	$\bar{y}=7{,}5$

$s_{xy} \approx 5{,}55$ $s_{xx} \approx 8{,}3$

$$\hat{\beta} = \frac{5{,}5}{8{,}3} \approx 0{,}67$$

Sorry, entspricht in diesem Datensatz zufällig r,
weil $s_x = s_y$ und damit $s_x^2 = s_x \cdot s_y$

$$\hat{\alpha} = 7{,}5 - 0{,}67 \cdot 5{,}5 \approx 3{,}82$$



$$f(x) = 3{,}82 + 0{,}67\,x$$

3.2.

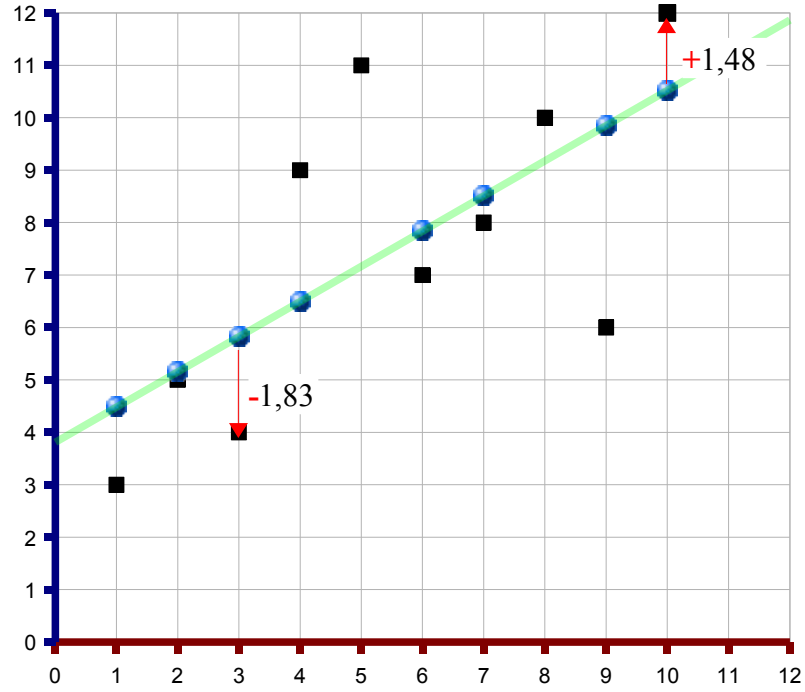
Berechnung der einzelnen Fehler

Residuen

$\hat{e}_i$

$\hat{e}_i = y_i - \hat{y}_i$ $\quad$ $\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$

x	1	2	3	4	5	6	7	8	9	10	
y	3	5	4	9	11	7	8	10	6	12	
$\hat{y}_i$	4,49	5,16	$3{,}82+0{,}67\cdot 3=$ 5,83	6,5	7,17	7,84	8,51	9,18	9,85	10,52	$\hat{y}_i + e_i = 0$
$\hat{e}_i$	-1,49	-0,16	$4-5{,}83=$ -1,83	2,5	3,83	-0,84	-4,51	0,82	-3,85	1,48	

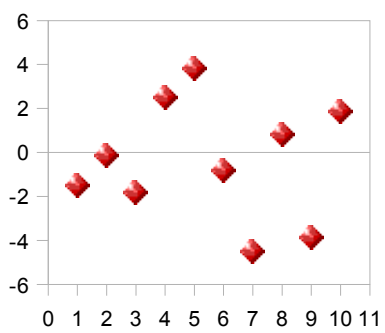


Streudiagramm der Residuen/Residualanalyse

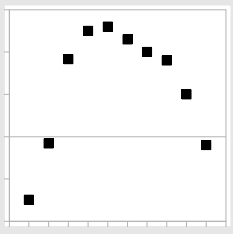
Residualplot

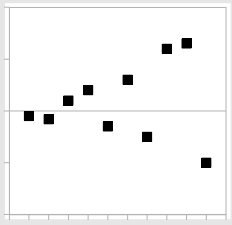
ideal:
Die Residuen streuen **unsystematisch** und **möglichst nahe bei 0**

x	1	2	3	4	5	6	7	8	9	10
$\hat{e}_i$	-1,49	-0,16	-1,83	2,5	3,83	-0,84	-4,51	0,82	-3,85	1,84



andere Residualplots weisen darauf hin, dass





der Datensatz nicht die Voraussetzungen für ein lineares Regressionsmodell erfüllt

alternative Plots (statt $x, \hat{e}$) : $\hat{y}, \hat{e}$ oder *(mit standardisierten Residuen)* $\hat{y}, \hat{d}$ ...

3.3.

Berechnung der unterschiedlichen Varianzen/ Varianzanteile

Streuungszerlegung		$SQT = SQE + SQR$
Sum of S**Q**uares **T**otal: Gesamtstreuung	Streuung der y-Werte um ihren Mittelwert $\bar{y}$	$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$
Sum of S**Q**uares **E**xplained: erklärte Streuung	Streuung der von der Regressionsgeraden erfassten $\hat{y}$ -Werte um den Mittelwert $\bar{y}$	$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
Sum of S**Q**uares **R**esidual: Residualstreuung / Reststreuung	Streuung der von der Regressionsgeraden nicht erfassten y -Werte um ihren entsprechenden Wert $\hat{y}$ auf der Regressionsgeraden *"Streuung der Fehler* $\hat{e}_i = y_i - \hat{y}_i$ *"*	$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
wäre im Beispiel ungefähr $82{,}3 = 37{,}1 + 45{,}2$ *(ziemlich großzügig passend gemacht ,*  *)*		

3.4.

Wie genau lassen sich mit dem Modell *(hier: einfache lineare Regression)* die y-Werte vorhersagen?

Bestimmtheitsmaß / Determinationskoeffizient

R^2

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2} = \frac{SQE}{SQT} \Rightarrow R^2 = r_{xy}^2$$

- Wertebereich **zwischen 0 und 1**

- je näher an 0, desto schlechter ist die Vorhersage

- im Fall $R^2 = 1$ liegen bereits die Originaldaten auf einer Geraden

$$R^2 = \frac{37,1}{82,3} \approx 0,45 \qquad \approx r^2 = 0,67^2$$

heißt: Das Modell eignet sich nicht so doll, um irgendwelche y-Werte vorauszusagen,

(z.B. "wie wäre der y-Wert zu x= 20?")

weil nur 45% der Gesamtstreuung von y durch die Regression erklärt werden

=

Die Varianz der y-Werte ist nur zu 45% auf die Varianz der x-Werte zurückzuführen.