Korrelation & Regression

Marcus Hudec

Notation

Den Ausgangspunkt bilden n Beobachtungspaare (x_i, y_i), die wir an n Merkmalsträgern erhoben haben und die wir als Datenpunkte in einem Streudiagramm (scattergram) visualisieren können.

Die Merkmale X und Y seien stetig (zumindest Intervallskalen-Niveau).

X	Y
x_1	y_1
x_2	y_2
...	...
x_i	y_i
...	...
x_n	y_n

Datensatz

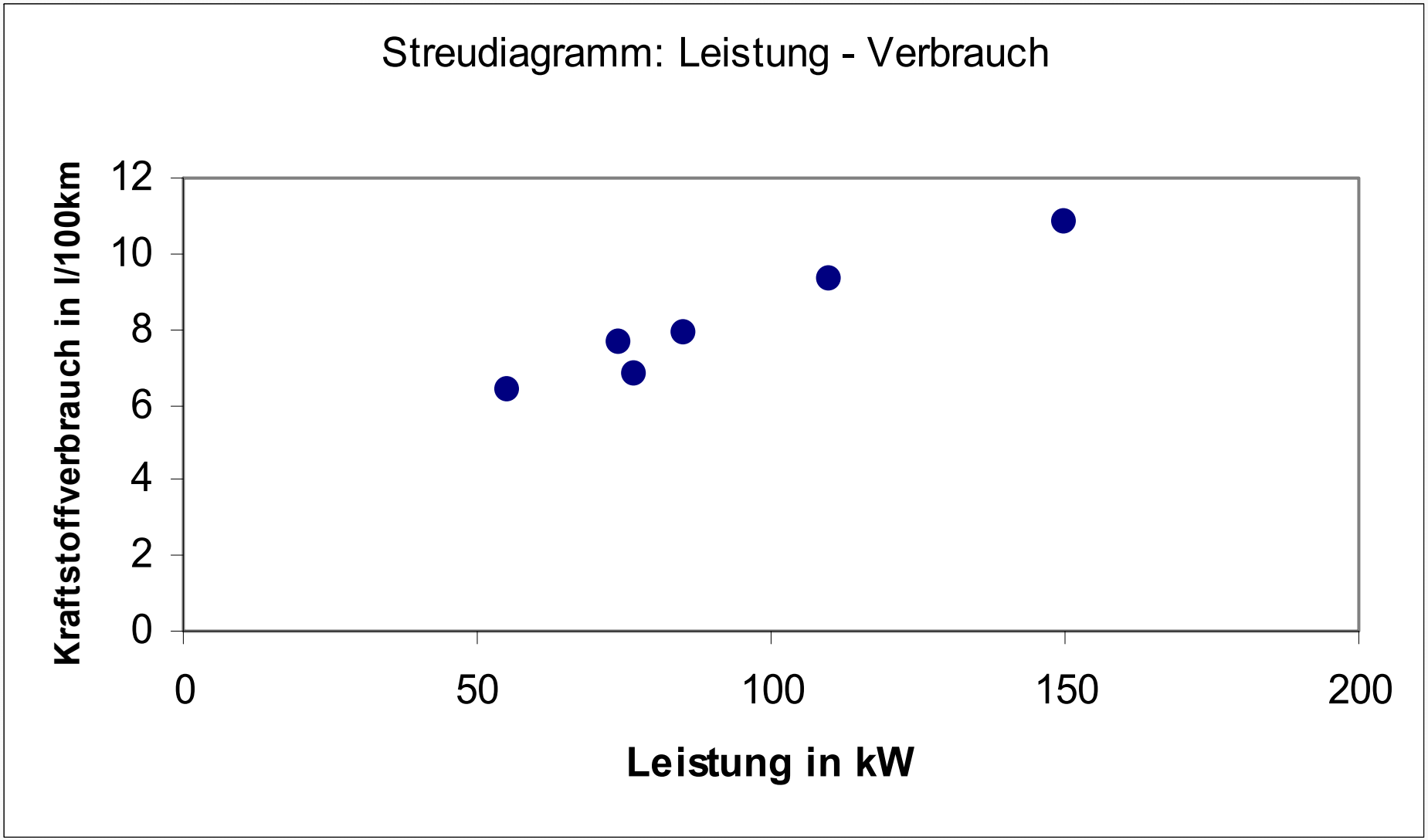
Beispiel

Leistung in kw und Kraftstoff-Verbrauch in l pro 100 km von sieben verschiedenen VW-Golf Benzinmotoren[1]

kw	l/100km
55	6,4
74	7,6
77	6,8
85	7,9
110	9,3
150	10,8

[1] Quelle: http://www.vw-online.de/golf/index_.htm

Streudiagramm (Scattergram)

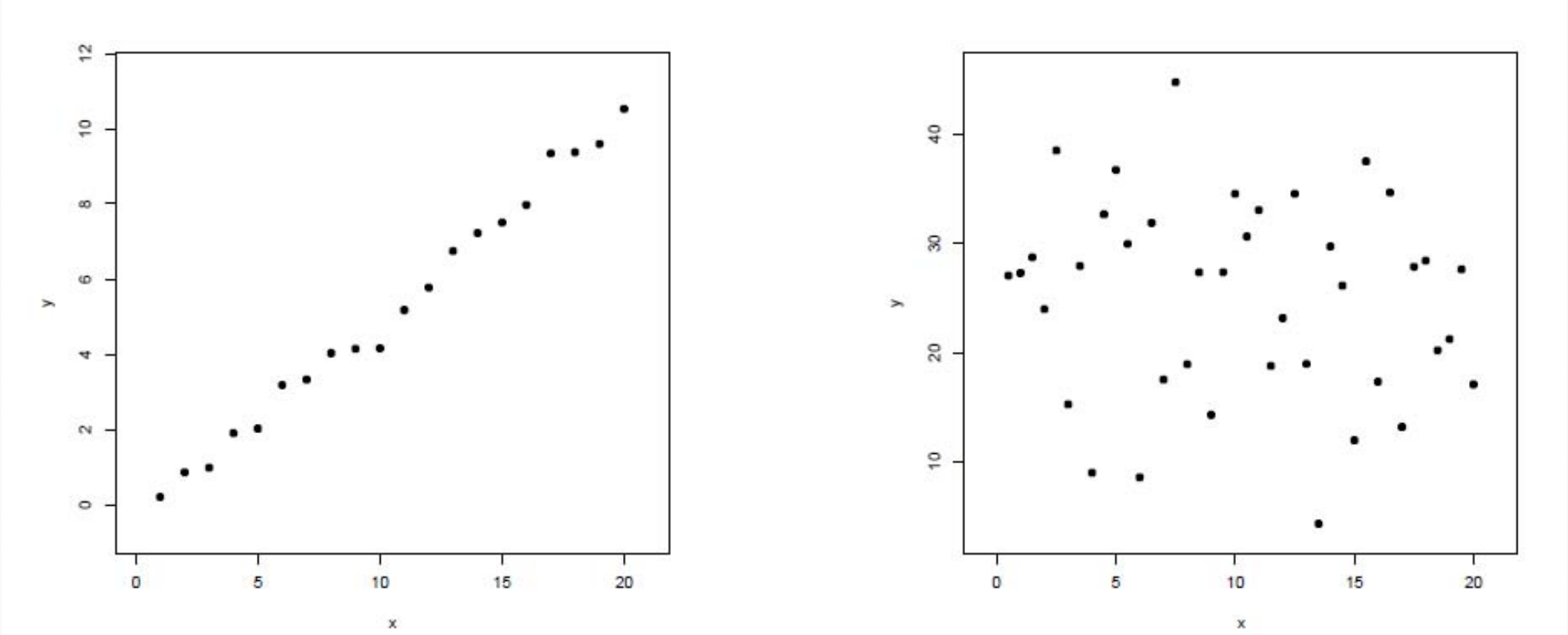


Streudiagramm

Ein Streudiagramm gibt einen Überblick über die gemeinsame Verteilung und liefert Antwort auf folgende Fragen:

- Liegt überhaupt ein Zusammenhang zwischen den beiden untersuchten Variablen vor?
- Kann dieser Zusammenhang in etwa durch eine lineare Funktion (Gerade) beschrieben werden?
- Wie stark ist Zusammenhang?
- Gibt es atypische Punkte (Ausreißer)?

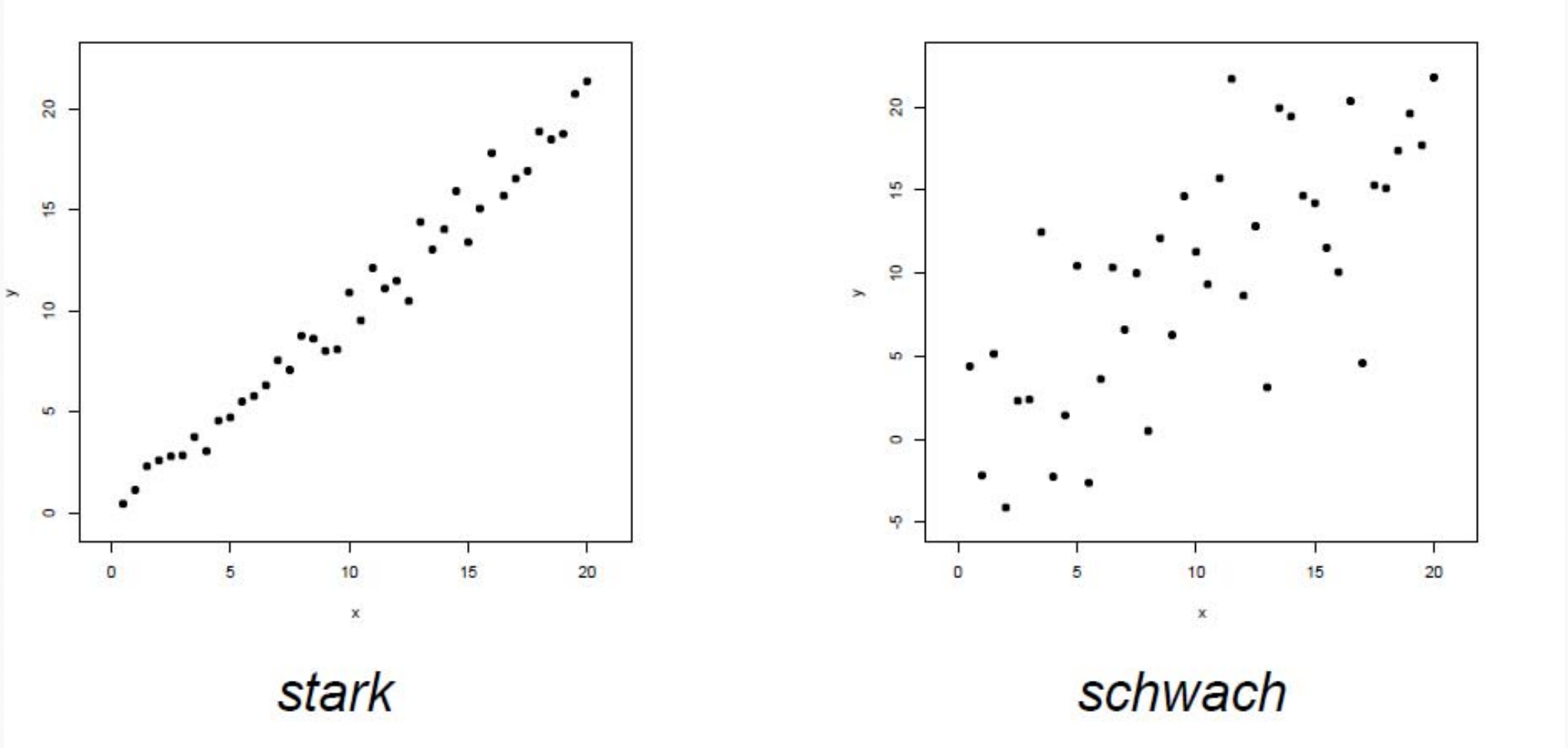
Beispiele



Linearer Zusammenhang

kein Zusammenhang

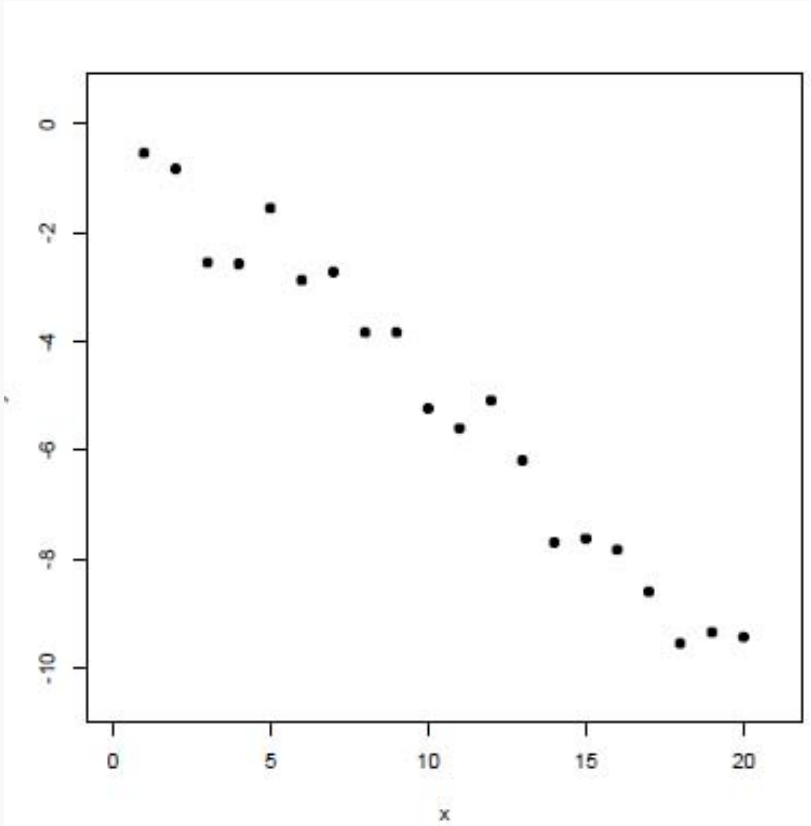
Beispiele



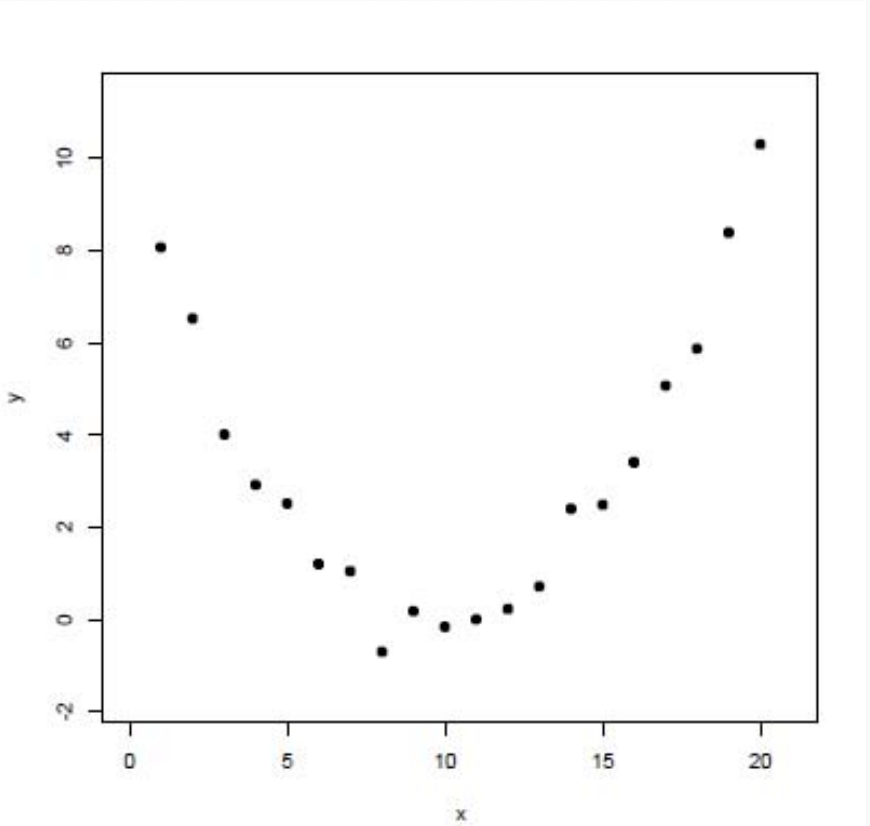
stark *schwach*

positiver Zusammenhang

Beispiele



Linear negativer Zusammenhang



Nichtlinearer Zusammenhang

Zusammenhang von 2 metrischen Merkmalen

- Wir betrachten den Fall der Messung des Zusammenhangs von 2 metrischen Variablen.
- Ziel ist es die Stärke und die Richtung des Zusammenhangs zwischen zwei Variablen X und Y mittels einer statistischen Maßzahl zu quantifizieren.
- Wir sprechen von einem positiven Zusammenhang, wenn Aussagen der Art: „Je größer X desto größer ist auch Y“ zutreffen
- Wir sprechen von einem negativen Zusammenhang, wenn Aussagen der Art: „Je größer X desto kleiner ist Y“ zutreffen

Kovarianz

Kovarianz:
Zusammenhangsmaß bei intervallskalierten Merkmalen, das sich unmittelbar aus der Varianz ableitet

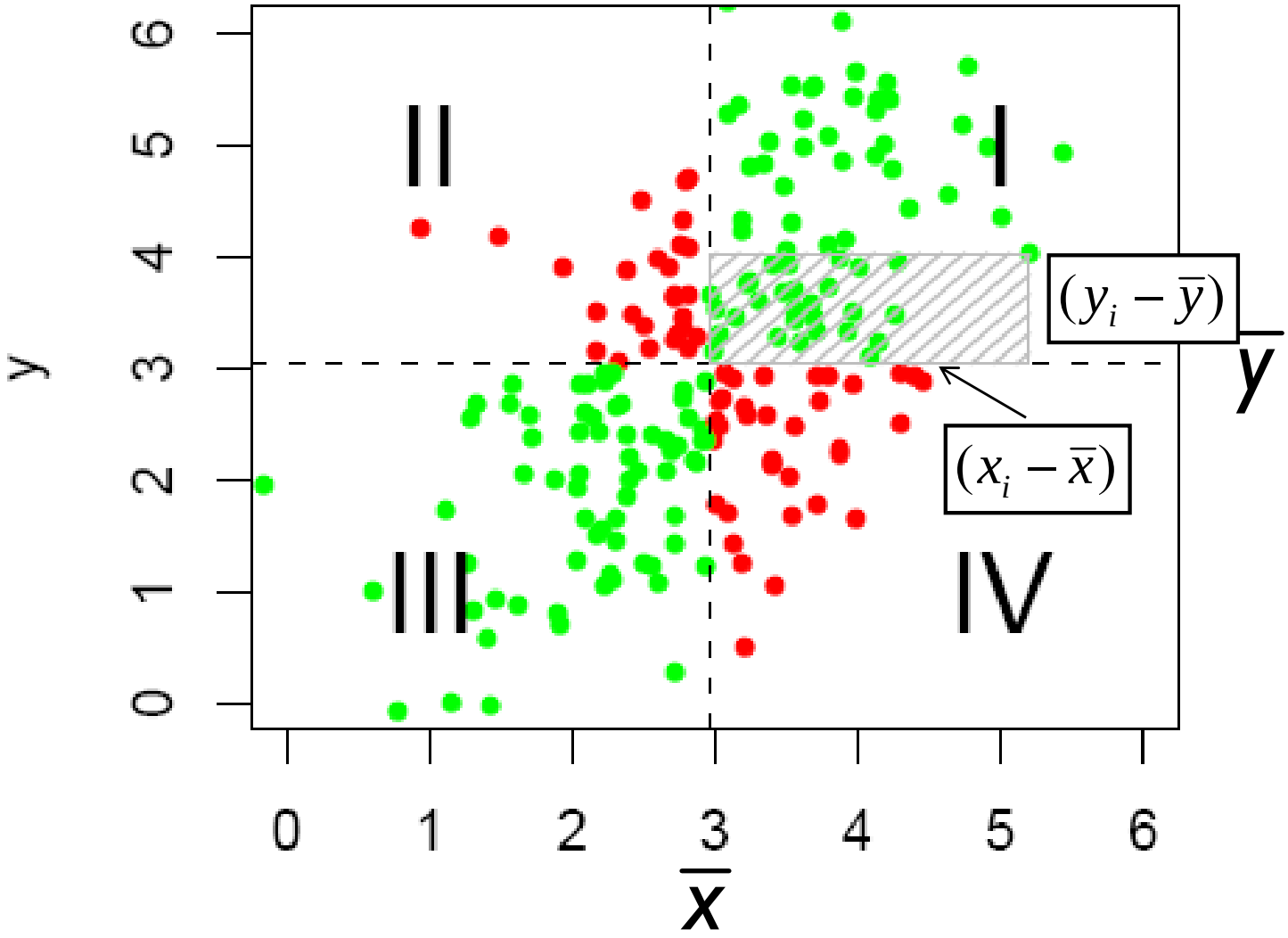
$$s_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i x_i - n\bar{x}\bar{x}\right)$$

Varianz von X

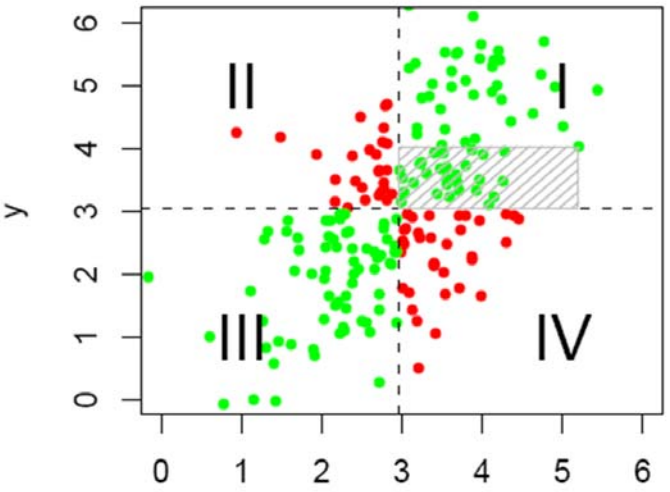
$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

Kovarianz von X und Y

Konzept der Kovarianz



Erklärung



- Jeder Beobachtungspunkt liefert einen Beitrag zur Summe.
- Grüne Punkte in den Quadranten I und III liefern positive Beiträge
- Rote Punkte in den Quadranten II und IV liefern negative Beiträge
- Die Größe des Beitrags entspricht der grau schraffierten Fläche

Nachteil:
Die Kovarianz ist nicht normiert und kann beliebige Werte aufweisen

Korrelationskoeffizient

Der Korrelationskoeffizient nach Pearson ist das wichtigste Maß für den Zusammenhang zwischen zwei metrischen Variablen X und Y und ergibt sich durch die Normierung der Kovarianz. Er ist ein Maß für die lineare Korrelation!

Alternative Bezeichnungen für dieses Maß in der Literatur:

- *Produkt-Moment-Korrelation*,
- *Bravais-Pearson-Korrelation*,
- *Linearer Korrelationskoeffizient*

Korrelationskoeffizient

Der Korrelationskoeffizient nach Pearson kann durch folgende äquivalente Formeln charakterisiert werden:

$$r_{xy} = corr_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\mathrm{cov}(XY)}{Std.Abw.(X) \times Std.Abw.(Y)} =$$

$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Hinweis:
wenn klar ist, um welche Korrelation es sich handelt wird oft auch nur r statt r_{XY} geschrieben.

Korrelationskoeffizient

Der Korrelationskoeffizient liegt stets zwischen -1 und +1.

Korrelationskoeffizient nahe -1:

Die Mehrzahl der Datenpunkte konzentrieren sich um eine Gerade mit negativer Steigung.

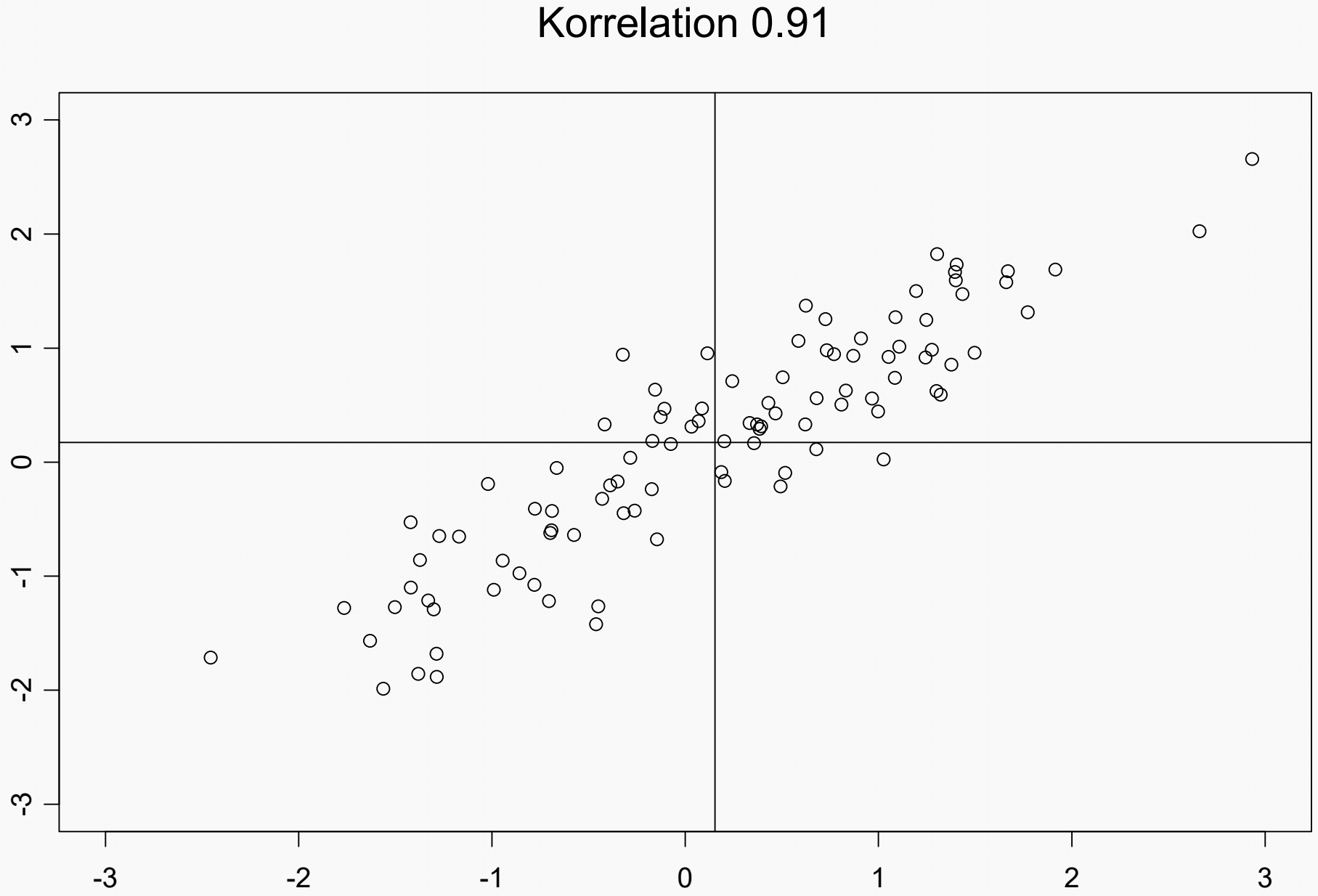
Korrelationskoeffizient ungefähr 0:

Die Datenpunkte sind entweder auf alle vier Quadranten ungefähr gleichmäßig verteilt oder sie liegen um eine Gerade die parallel zu einer Achse verläuft.

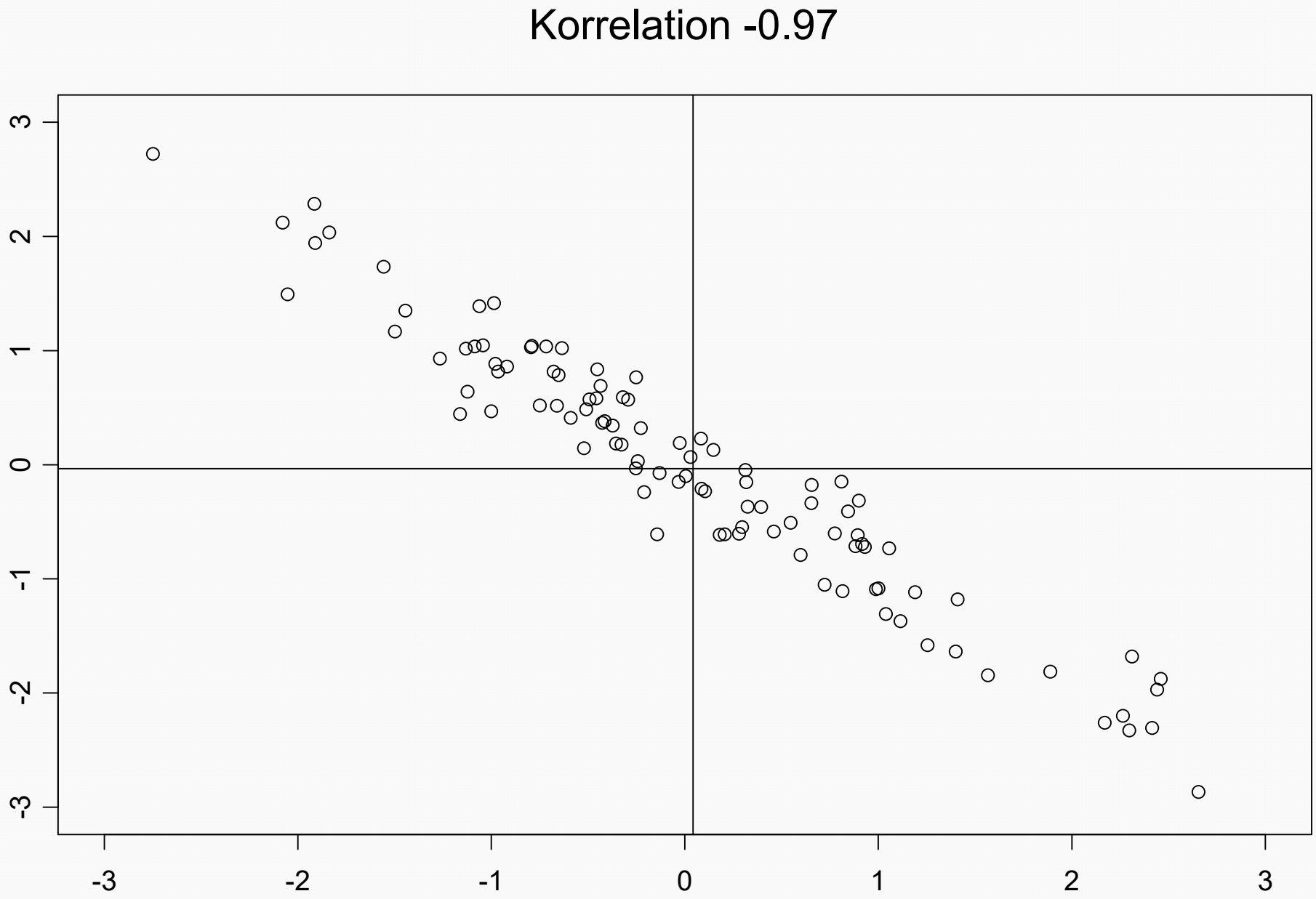
Korrelationskoeffizient nahe +1:

Die Mehrzahl der Datenpunkte konzentrieren sich um eine Gerade mit positiver Steigung.

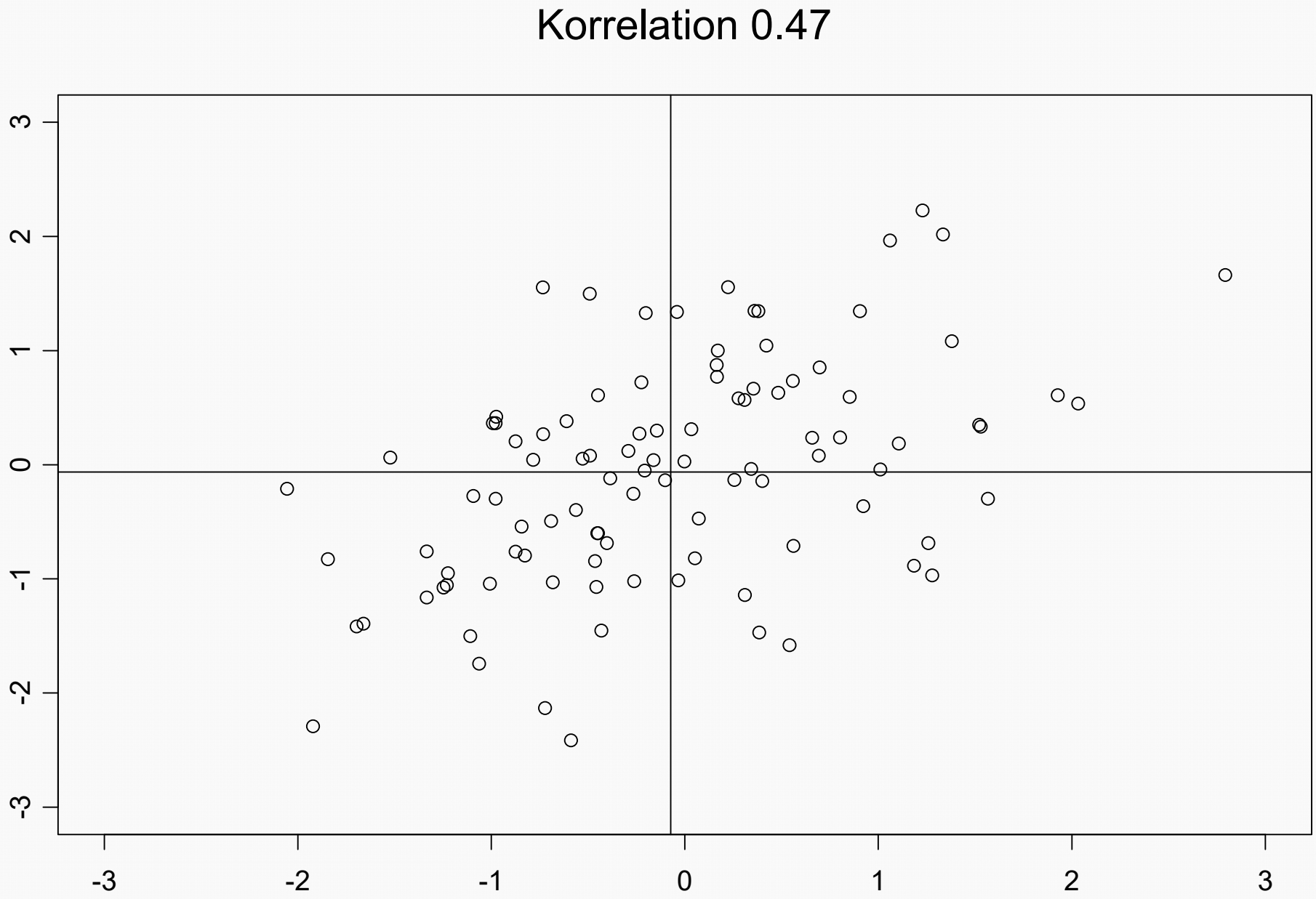
Hohe positive Korrelation



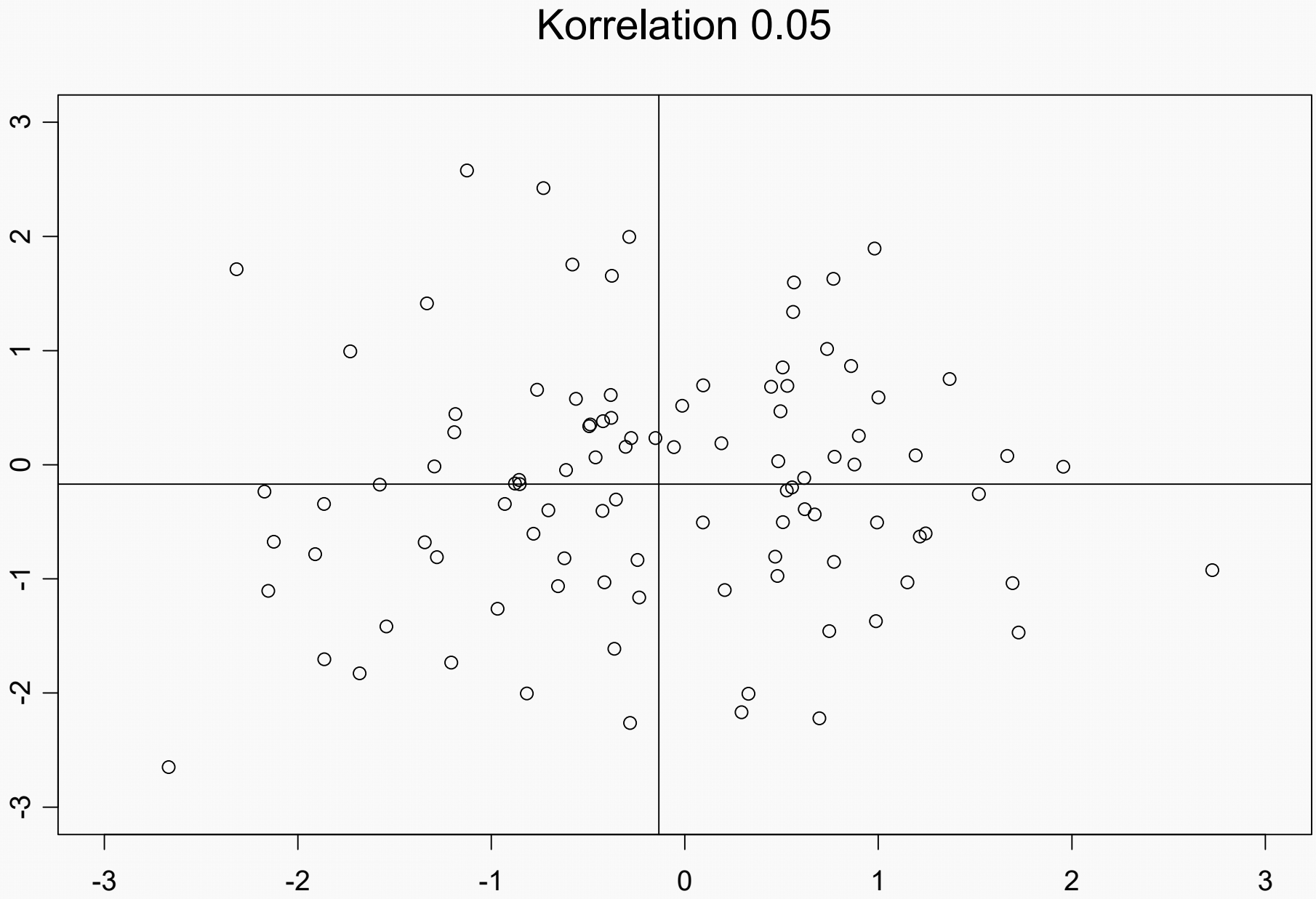
Hohe negative Korrelation



Mittlere positive Korrelation



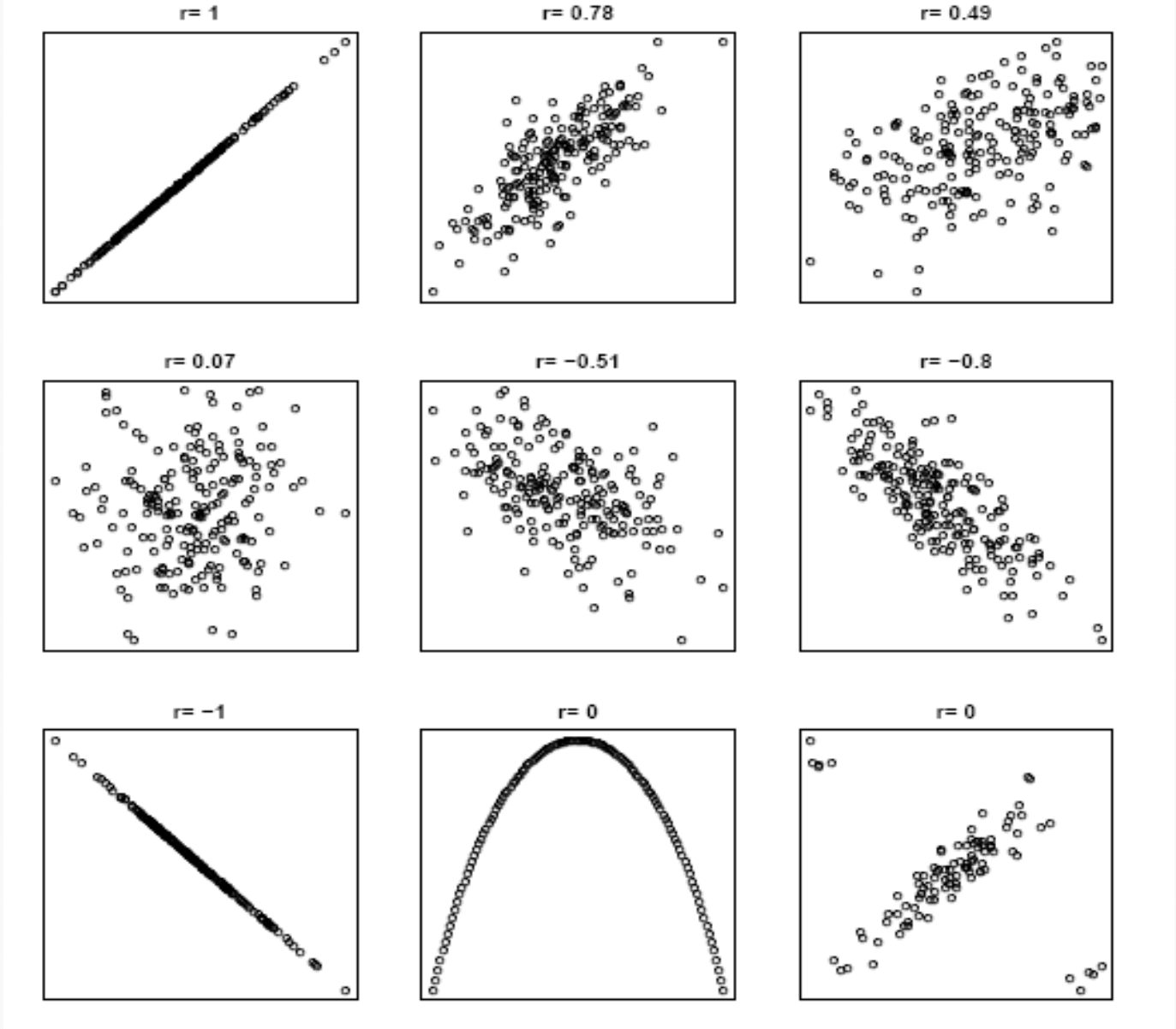
Korrelation nahe 0



Was ist eine starke Korrelation?

- **Vorschlag von Cohen**:
- |r| ~ 0,1 schwacher Zusammenhang
- |r| ~ 0,3 mittlerer Zusammenhang
- |r| ~ 0,5 starker Zusammenhang
- Ist r deutlich größer als 0,5 spricht man von einem sehr starken Zusammenhang
- Beachte: Die Relevanz des Wertes ist immer auch in Abhängigkeit vom Stichprobenumfang n zu sehen.

Verschiedene Szenarien



Test auf Signifikanz der Korrelation

Will man Hypothesen der Form

H_0: corr =0 versus H_a: corr≠0 (zweiseitig)

bzw.

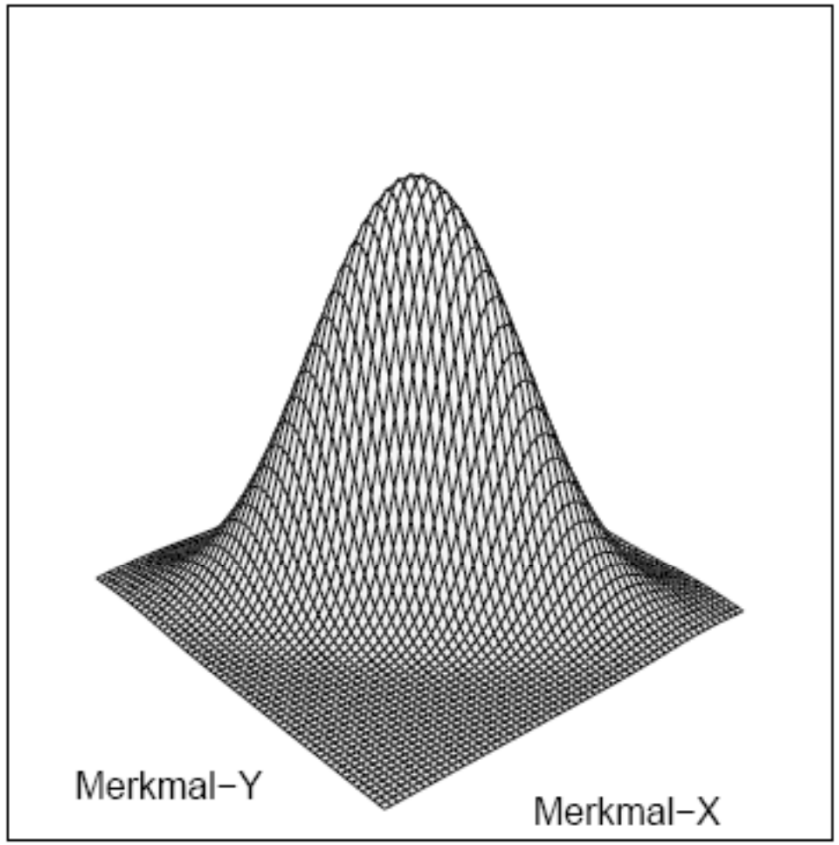
H_0: corr<0 versus H_a: corr>0 (einseitig)

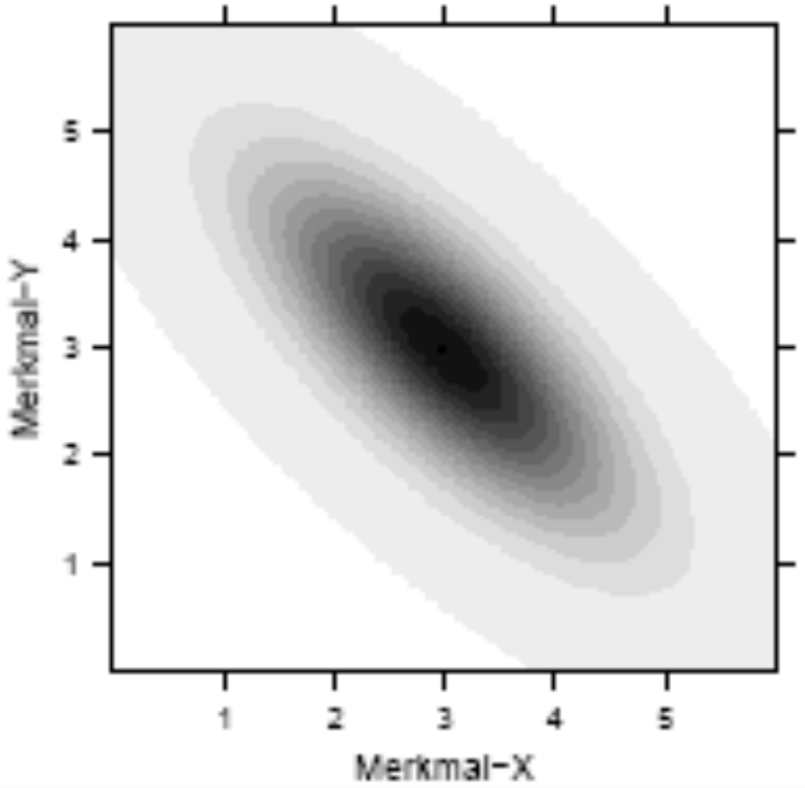
testen, so kann dies unter der Annahme einer 2-dimensionalen Normalverteilung mit folgender Statistik erfolgen:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{mit} \quad n-2 \quad \text{Freiheitsgraden}$$

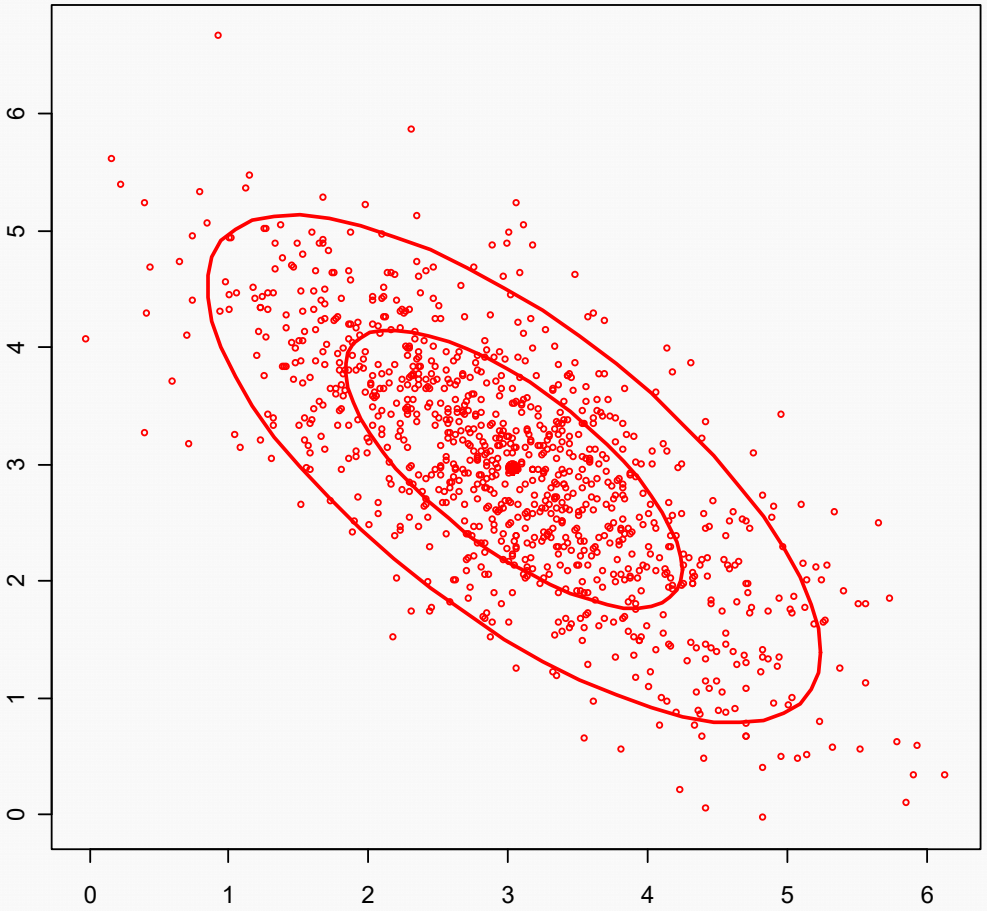
Diese Teststatistik ist unter der Nullhypothese t verteilt mit n-2 Freiheitsgraden.

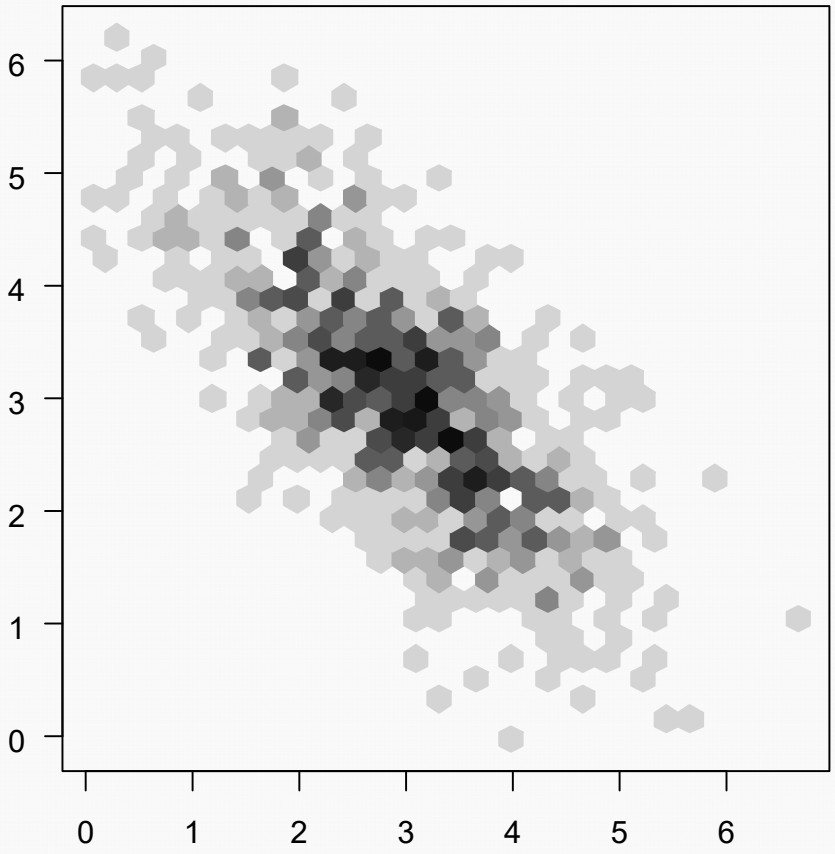
Bivariate Normalverteilung

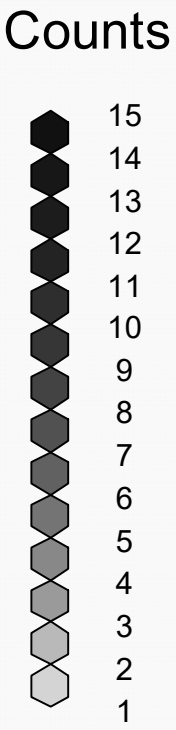




Simulation aus einer 2-dimensionalen Normalverteilung







Beispiel

i	X	Y	X^2	XY	Y^2
1	65	68	4225	4420	4624
2	63	66	3969	4158	4356
3	67	68	4489	4556	4624
4	64	65	4096	4160	4225
5	68	69	4624	4692	4761
6	62	66	3844	4092	4356
7	70	68	4900	4760	4624
8	66	65	4356	4290	4225
9	68	71	4624	4828	5041
10	67	67	4489	4489	4489
11	69	68	4761	4692	4624
12	71	70	5041	4970	4900
Summe	800	811	53418	54107	54849

Korrelationen

		X	Y
X	Korrelation nach Pearson	1	,703*
	Signifikanz (2-seitig)		,011
	N	12	12
Y	Korrelation nach Pearson	,703*	1
	Signifikanz (2-seitig)	,011	
	N	12	12

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Wir wollen die Nullhypothese testen, ob die Merkmale X und Y unkorreliert sind.

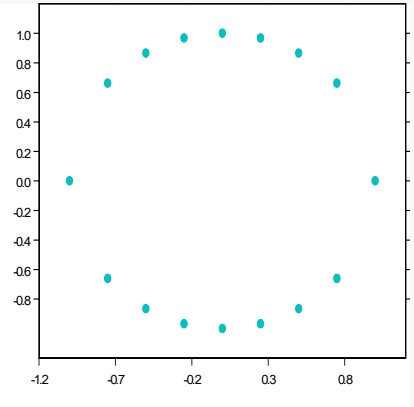
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Kovarianz	Sxy	484			
Varianz X	Sxx	1016			
Varianz Y	Syy	467			
Korrelation	Rxy	0,70			
Teststatistik	Zähler	2,22			
	Nenner	0,71			
	t	3,12			
Tabellenwert	$t_{n-2;0,975}$	2,23	==> Ho ablehnen		

```
> 2*(1-pt(3.12, 10))
[1] 0.01087398
```

Unabhängigkeit und Kausalität

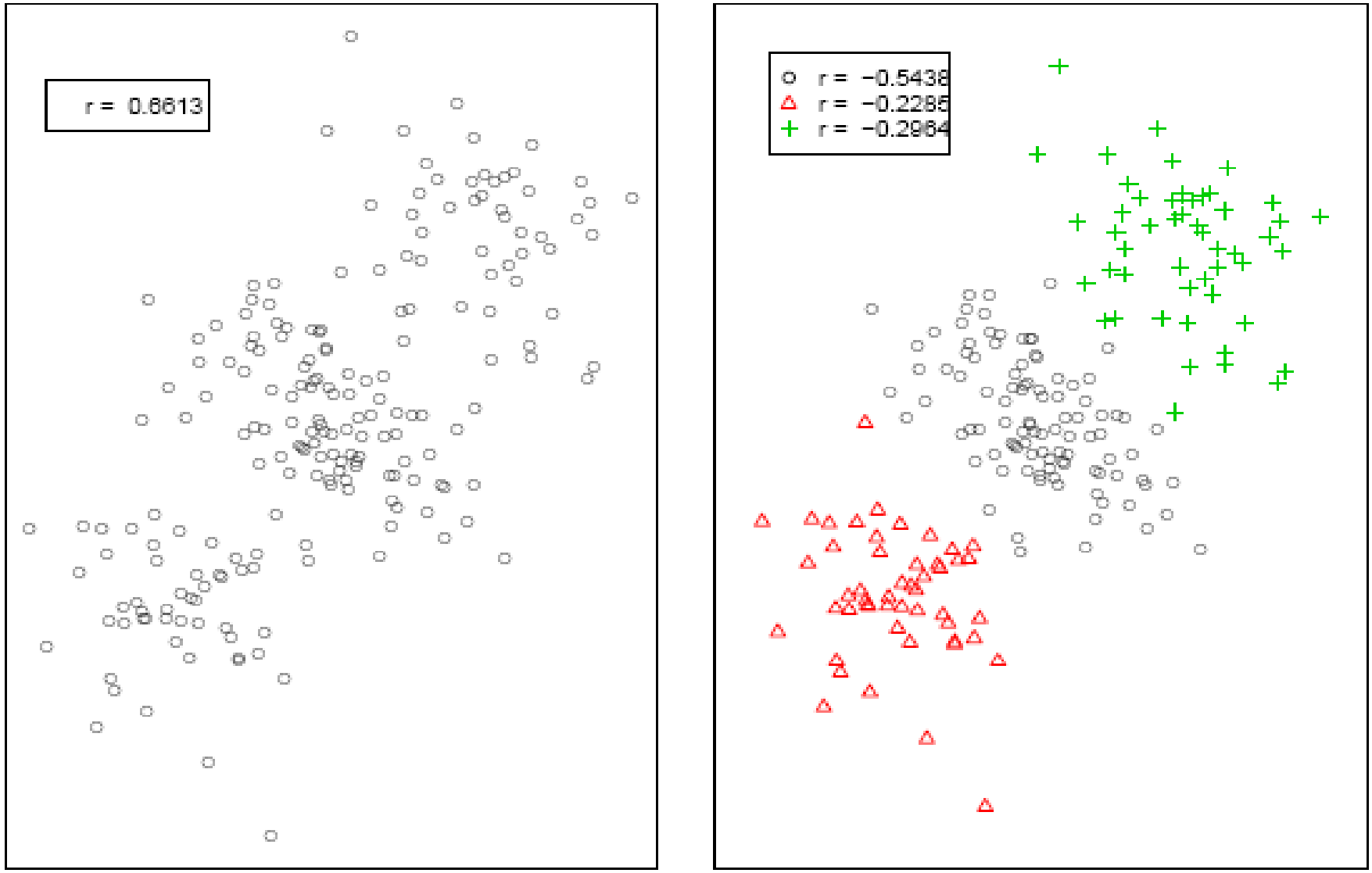
- Sind zwei Variablen statistisch unabhängig, so folgt daraus, dass der Korrelationskoeffizient den Wert 0 annimmt.
- Umgekehrt kann aus einer Korrelation von Nahe Null nicht auf Unabhängigkeit geschlossen werden, da die Korrelation nur den linearen Zusammenhang misst.



Die Punkte im linken Beispiel haben Korrelation null!

- Keinesfalls darf Korrelation mit Kausalität gleichgesetzt werden. Problem: Scheinkorrelation

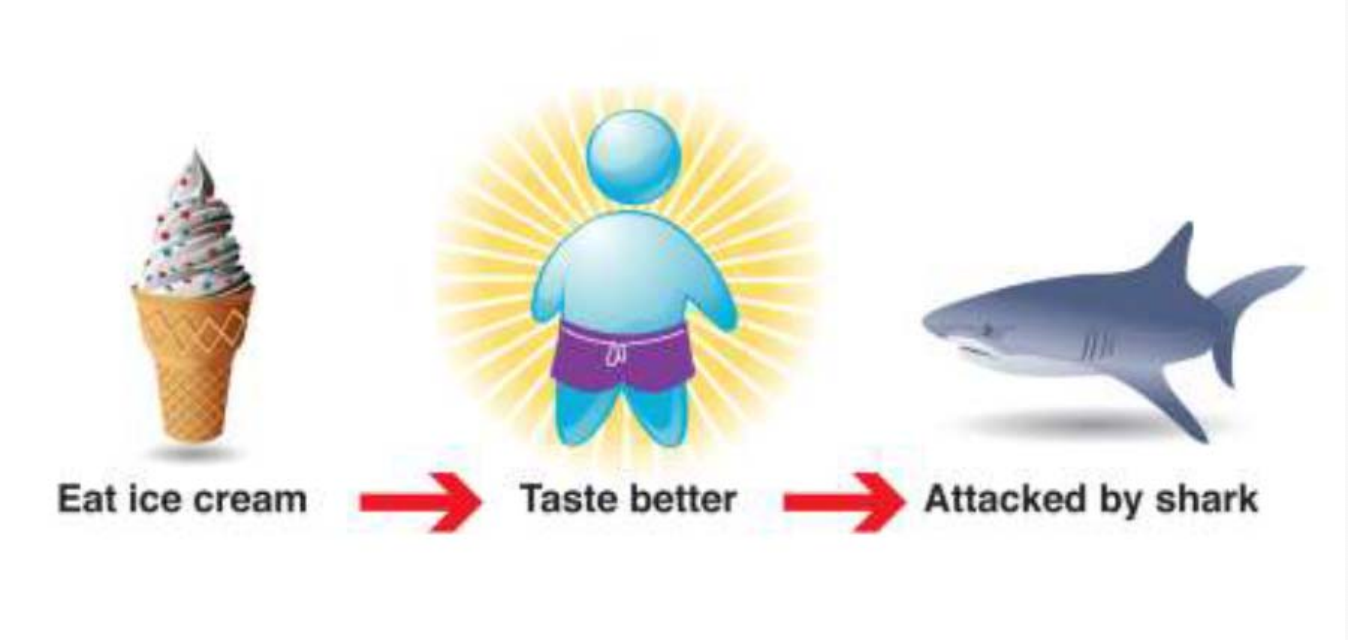
Simpsons Paradoxon (heterogene Gruppen)



Kausalität

- **Kausalität** bezeichnet die Beziehung zwischen **Ursache** und **Wirkung**, wobei die Ursache ein Sachverhalt ist, der einen bestimmten anderen Sachverhalt (Wirkung) als Folge herbeiführt. Kausalität weist eine feste Richtung auf, die immer von der Ursache ausgeht, auf der die Wirkung folgt.
- Korrelation ist ungerichtet (beachte Symmetrie in der Formel)
- Korrelation kann auch über Drittvariablen entstehen

Correlation vs. Causality

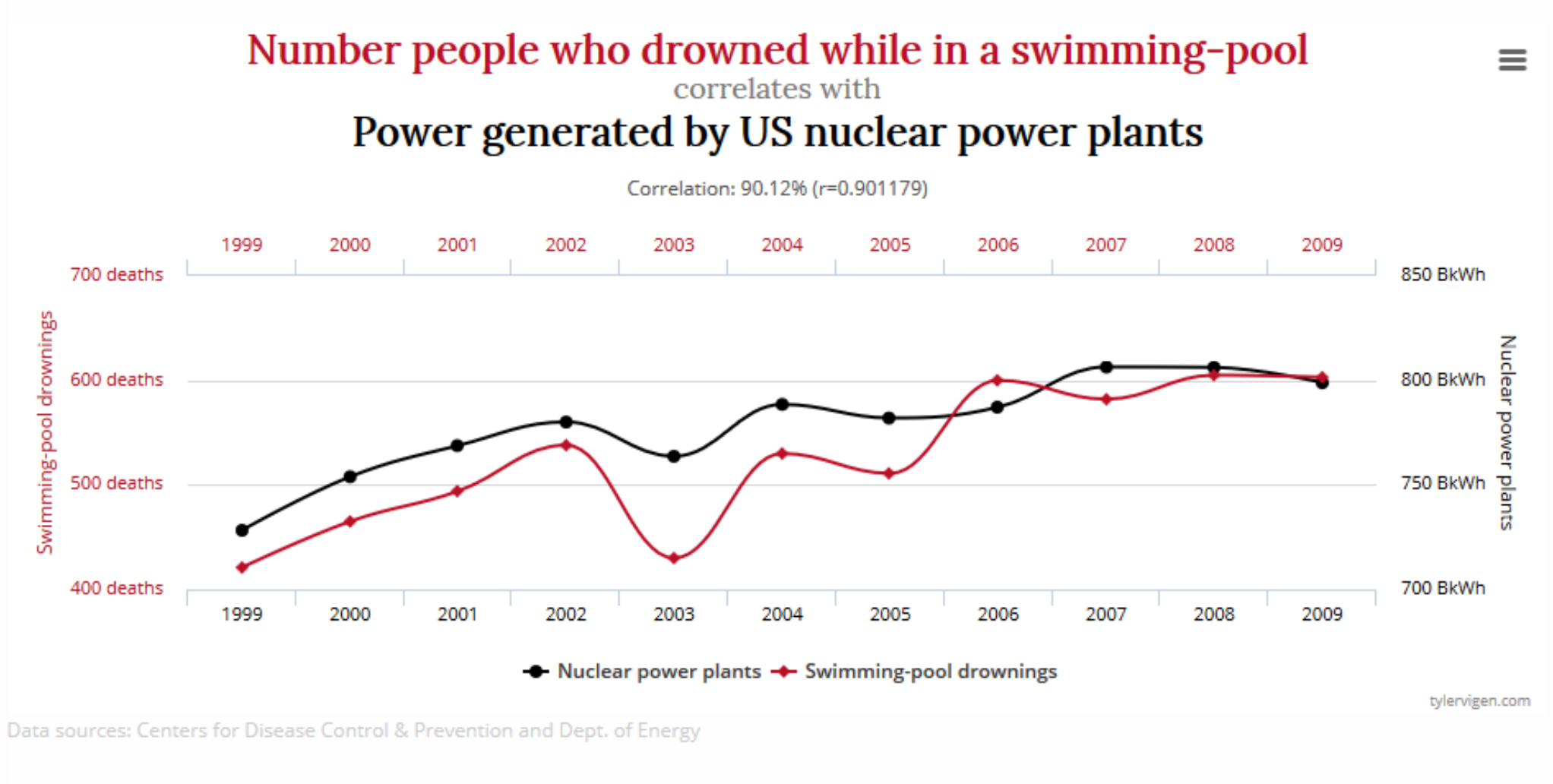


Empirische Daten zeigen, dass der Verzehr von Speiseeis das Risiko von einem Haifisch attackiert zu werden erhöht!

Quelle: Eric Siegel. Predictive Analytics: Delivering on the Promise of Big Data. IBM Government Analytics Forum, May2014

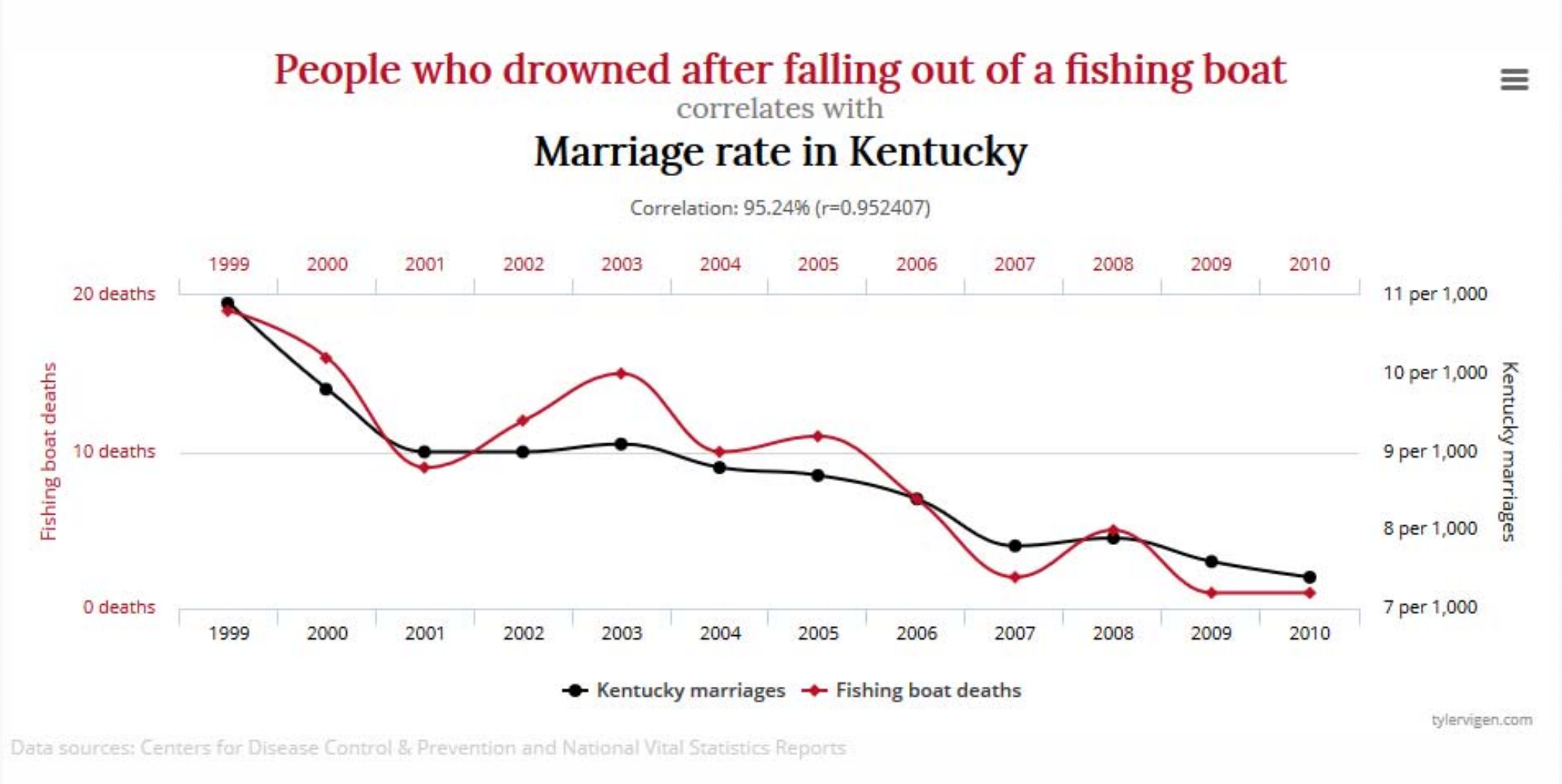
Scheinkorrelation

Nonsens-Korrelationen (1)



http://tylervigen.com/spurious-correlations

Nonsens-Korrelationen (2)



http://tylervigen.com/spurious-correlations

Korrelation bei ordinalen Daten

Für ordinalskalierte Variablen, eignet sich der **Rangkorrelationskoeffizient nach Spearman**.

Idee:

Verwende bei der Berechnung des Korrelations-koeffizienten nicht den Wert der Beobachtungen X und Y sondern den Rang, den diese Beobachtung aufgrund des Wertes bei einer Sortierung nach X bzw. Y einnimmt.

x_i, y_i … gemessene Werte für die i-te Beobachtung

Rx_i, ….. Rang, den die i-te Beobachtung bei Ordnung nach X einnimmt.

Ry_i, ….. Rang, den die i-te Beobachtung bei Ordnung nach Y einnimmt.

Rang-Korrelation nach Spearman

$$r_S = \frac{\sum_{i=1}^{n}(Rx_i - \bar{R}x)(Ry_i - \bar{R}y)}{\sqrt{\sum_{i=1}^{n}(Rx_i - \bar{R}x)^2 \sum_{i=1}^{n}(Ry_i - \bar{R}y)^2}}$$

$$D_i = Rx_i - Ry_i$$

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

Die Formel links ergibt sich als Vereinfachung der Formel für den Korrelationskoeffizienten, wenn man mit den Rängen arbeitet

Korrelation bei ordinalen Daten

Rang-Korrelation nach Spearman

Ist eine sogenannte nichtparametrische (robuste) Methode, da Inferenzaussagen mit weniger Modellvoraussetzungen möglich sind und die Ergebnisse nicht so stark von einzelnen Extremwerten beeinflusst werden. Das ist in etwa vergleichbar mit der Diskussion Median versus arithmetisches Mittel

Voraussetzungen:

Der aus der Punktwolke vermutete Zusammenhang muss nur **monoton** sein (monoton wachsend oder fallend). Es ist kein linearer Zusammenhang notwendig.

Beide Merkmale sind mindestens **ordinal skaliert**.

Eigenschaften:

Interpretation analog zum Korrelationskoeffizienten nach Pearson;

Sein Wert wird weniger stark von Ausreißern beeinflusst als der des Korrelationskoeffizienten nach Bravais und Pearson

Beispiel

Rangreihung von 5 Universitäten durch 2 Gutachter

Universität	Gutachter 1	Gutachter 2	d_i^2
A	1	3	4
B	2	1	1
C	3	2	1
D	4	5	1
E	5	4	1
			8

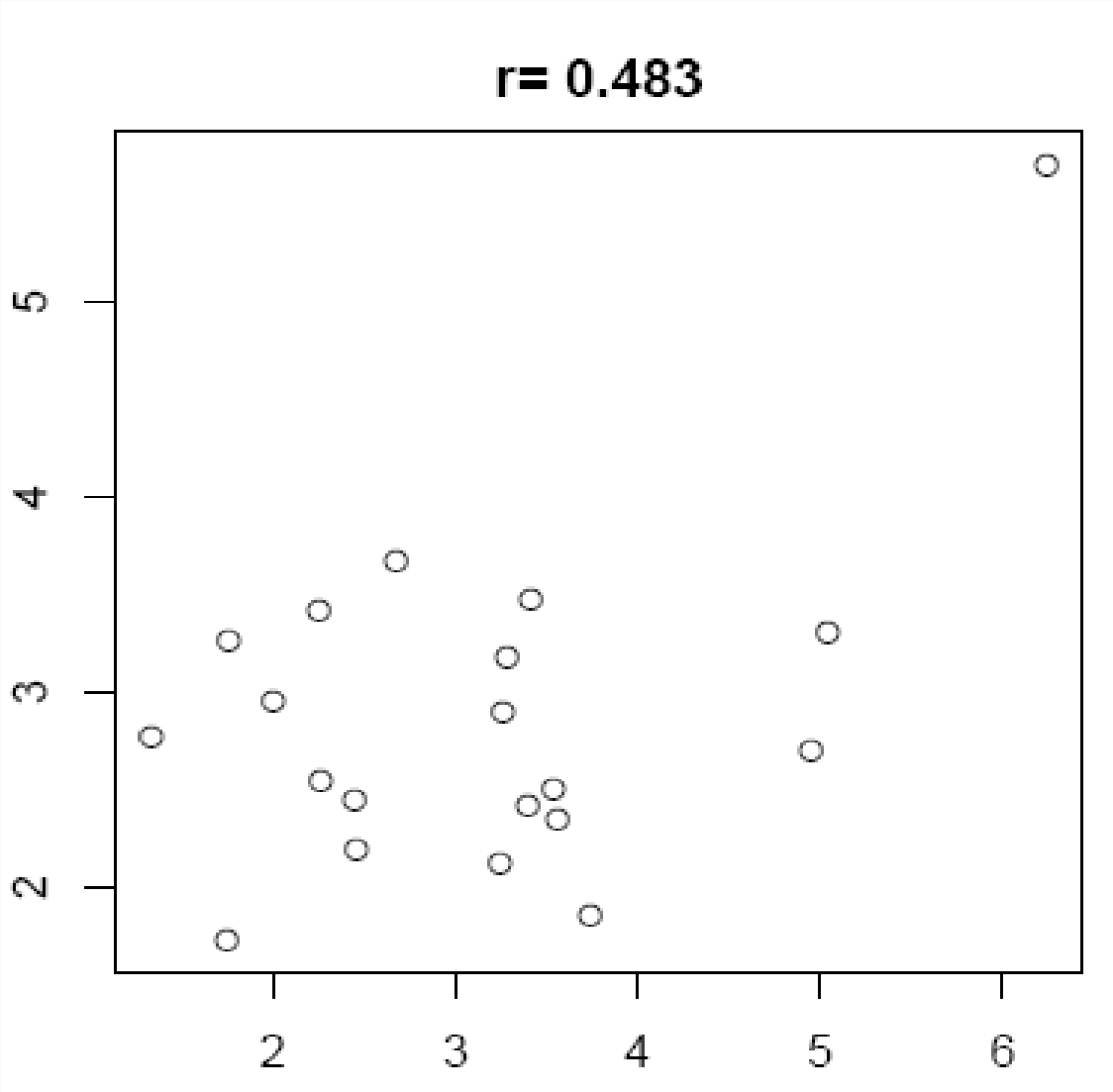
Rangkorrelation nach Spearman = 0,6

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

Rangkorrelation nach Spearman

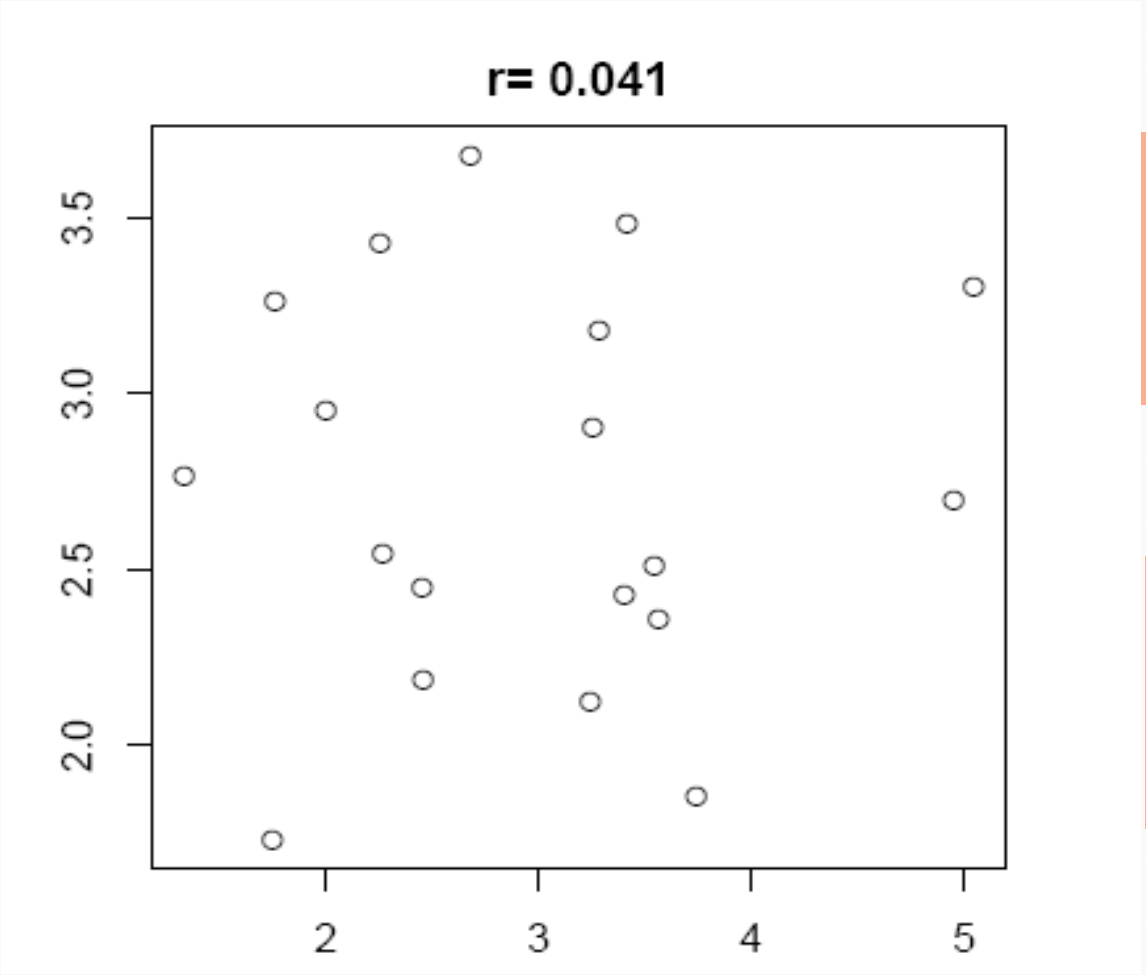
- Vorteile:
  - Bereits anwendbar auf zumindest ordinal-skalierte Daten
  - Keine Annahme, dass die Beziehung zwischen den Variablen linear ist.
  - Der Rangkorrelationskoeffizient ist robust gegenüber Ausreißern.
  - Invariant gegenüber monotonen Transformationen
- Nachteile:
  - Informationsverlust bei Vorliegen stetiger Merkmale
  - Insbesondere bei normalverteilten Daten resultiert daraus ein Genauigkeitsverlust

Vertrauen Sie dieser Korrelation?



Berechnung der Korrelation nach Bravais & Pearson

Elimination des extremen Datenpunkts



Berechnung der Korrelation nach Bravais & Pearson

Praktisch keine Korrelation in den übrigen Daten !!

Anwendung der Rangkorrelation

X	Y	Rang-X	Rang-Y	D
1.332	2.771	1	11	-10
1.748	1.730	2	1	1
1.758	3.266	3	15	-12
2.003	2.953	4	13	-9
2.252	3.429	5	17	-12
2.265	2.546	6	9	-3
2.449	2.452	7	7	0
2.460	2.191	8	4	4
2.679	3.682	9	19	-10
3.240	2.127	10	3	7
3.257	2.906	11	12	-1
3.285	3.183	12	14	-2
3.401	2.427	13	6	7
3.415	3.484	14	18	-4
3.541	2.512	15	8	7
3.563	2.358	16	5	11
3.739	1.856	17	2	15
4.952	2.700	18	10	8
5.042	3.309	19	16	3
6.245	5.716	20	20	0

Berechnung der Korrelation nach Spearman ergibt für das vollständige Sample:
r_s=0,1113

Durch die Reduktion der Skalierung erfolgt implizit eine schwächere Gewichtung extremer Beobachtungen

Nachteil: Informationsverlust
Vorteil: Robust gegenüber Datenfehlern

Vergleichbar mit der Diskussion Median versus arithmetisches Mittel

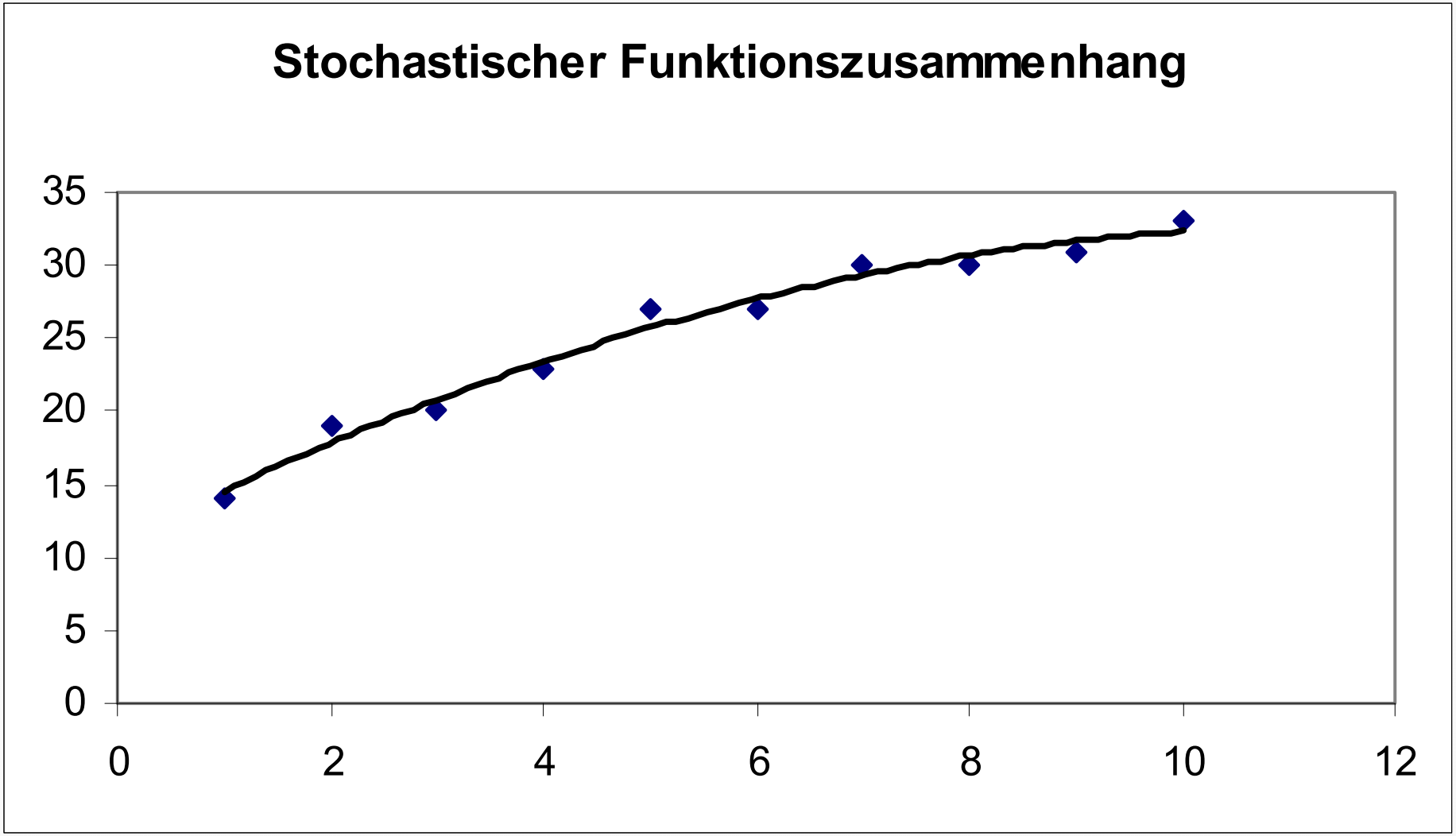
Grundmodell der einfachen Regression

- Zielgröße (abhängige Variable; Regressand) Y
- Einflussgröße (unabhängige Variable; Regressor) X
- Im Beispiel:
  - Y ... Kraftstoffverbrauch
  - X ... Leistung
- Annahme:

  Es besteht ein funktionaler Zusammenhang zwischen den beiden Merkmalen: $Y = f(X)$

- Die Regressionsanalyse ist ein Instrument zur Untersuchung eines **funktionalen Zusammenhangs** zwischen zwei Merkmalen.
- Im Unterschied zur Korrelationsanalyse handelt es sich also um ein gerichtetes Modell
- Dabei handelt es sich nicht um eine exakte Funktion im streng mathematischen Sinne
- Aufgrund von Messfehlern und Zufallseinflüssen werden die einzelne Messungen nicht idealtypisch auf dem Funktionsgraphen liegen, sondern zufällig abweichen
- Wir erweitern unser Modell daher um einen Fehlerterm (zufällige Komponente) e, wie folgt:

$$Y = f(X) + e$$



Lineare Regression

- Die einfachste Form eines funktionalen Zusammenhanges stellt eine lineare Funktion dar
- Modellvorstellung: der Zusammenhang zwischen X und Y kann (zumindest stückweise) durch eine Grade mit einem additiven Fehlerterm beschrieben werden:

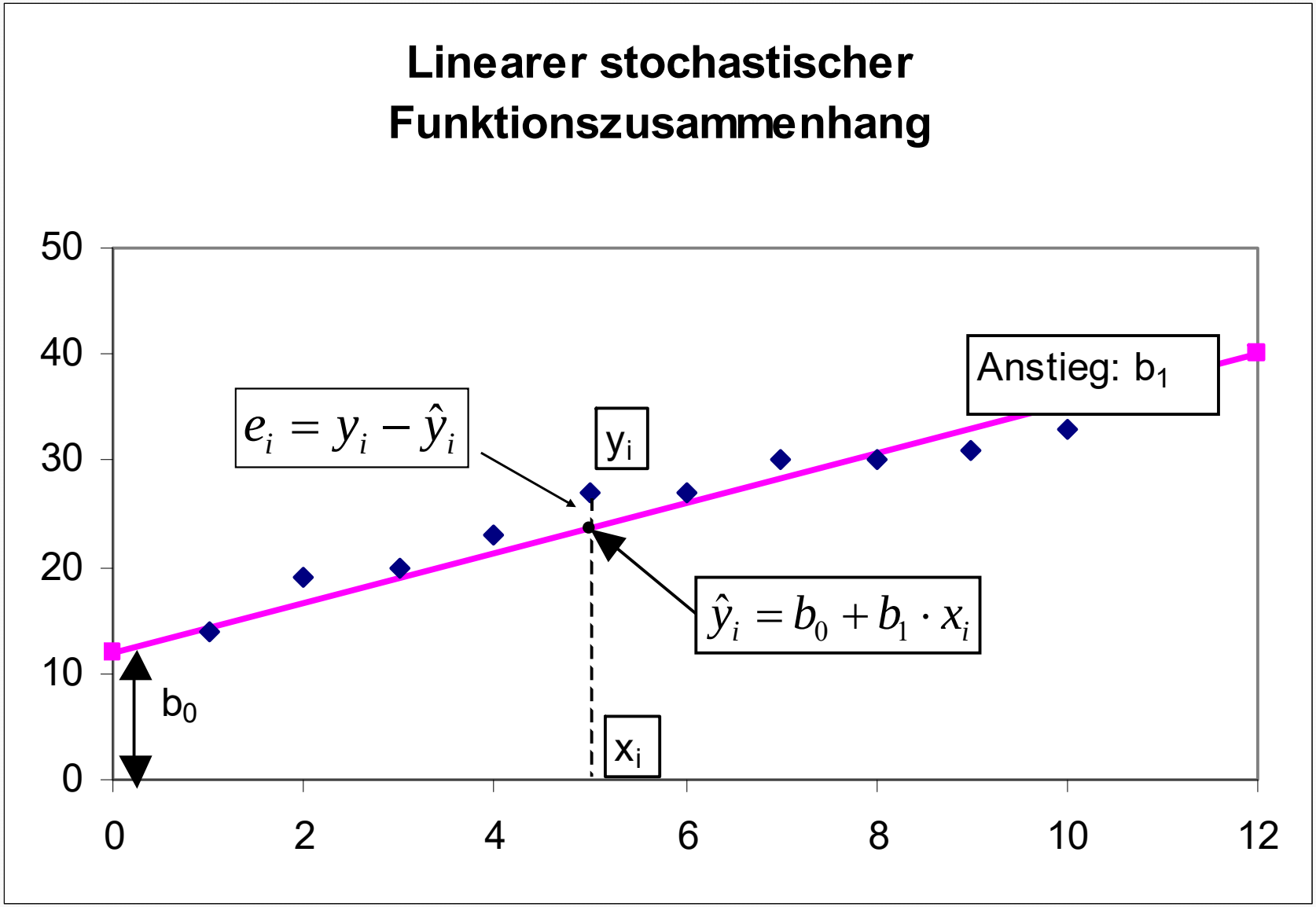
$$Y = b_0 + b_1 X + e$$

b_0 ... Abstand der Gerade vom Ursprung auf der Ordinate

b_1 ... Steigung der Gerade

e … zufälliger Fehler

Notation



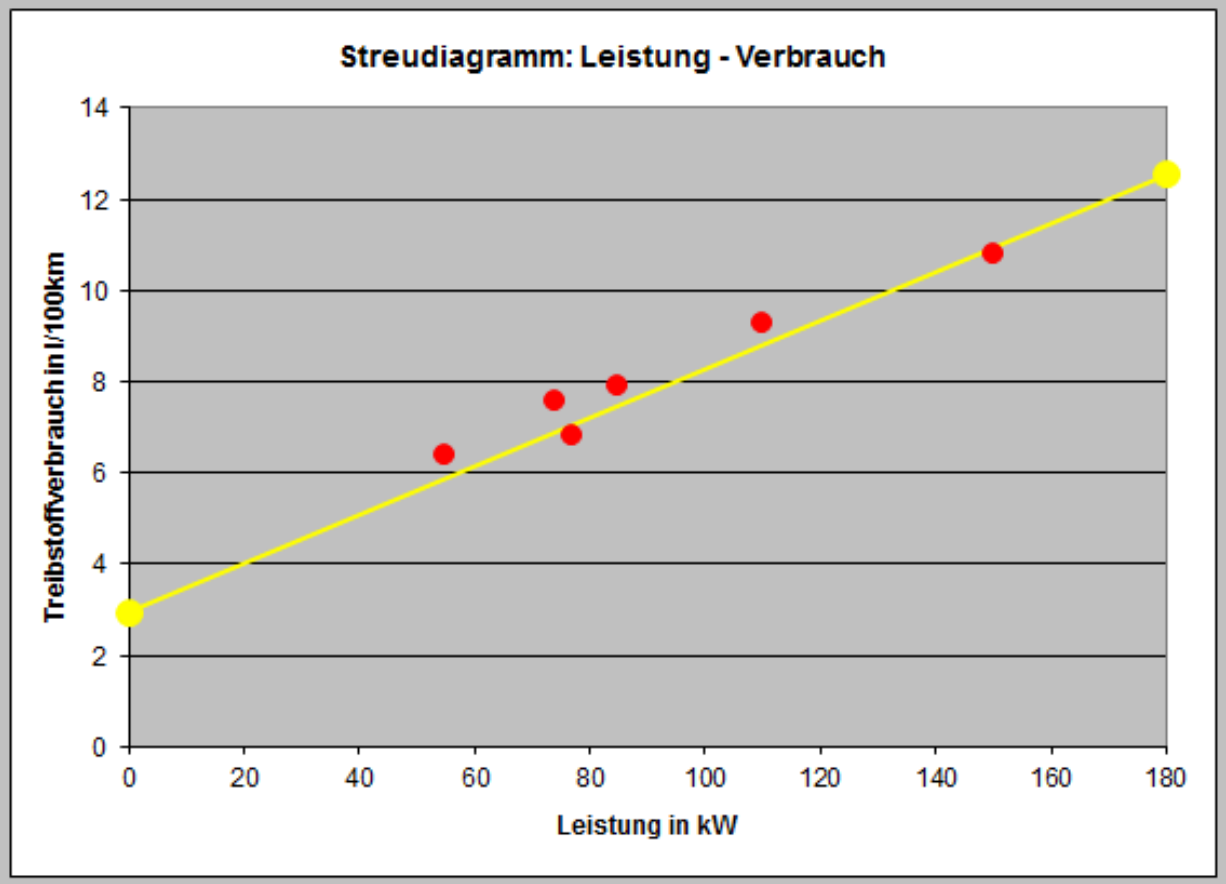
Intuitive Anpassung

Versuche, die gelbe Gerade nach Augenmass so gut als möglich den roten Datenpunkten anzupassen!

Mit Hilfe der beiden Schieberegler können die beiden gelben Ankerpunkte der Geraden nach oben und unten verschoben werden.

Beachte, wie sich die beiden Parameter der Geraden (Abstand und Steigung) verändern!

Wenn Du glaubst, den Zusammenhang gut beschrieben zu haben, kannst Du die (nach dem Kleinst-Quadrate Prinzip optimale) Regressionsgerade einblenden.



Abstand: $b_0 = 2{,}95$

Steigung: $b_1 = 0{,}05$

Regressionsgerade einblenden

Regressionsgerade ausblenden

links rechts

Bestimmung der Regressionsgeraden

Idee: Gerade so legen, daß die Summe der QUADRATE aller Abweichungen minimal wird.

Kleinst-Quadrate-Prinzip

Die optimale Regressionsgerade ergibt sich dann durch Lösung

folgender Optimierung:

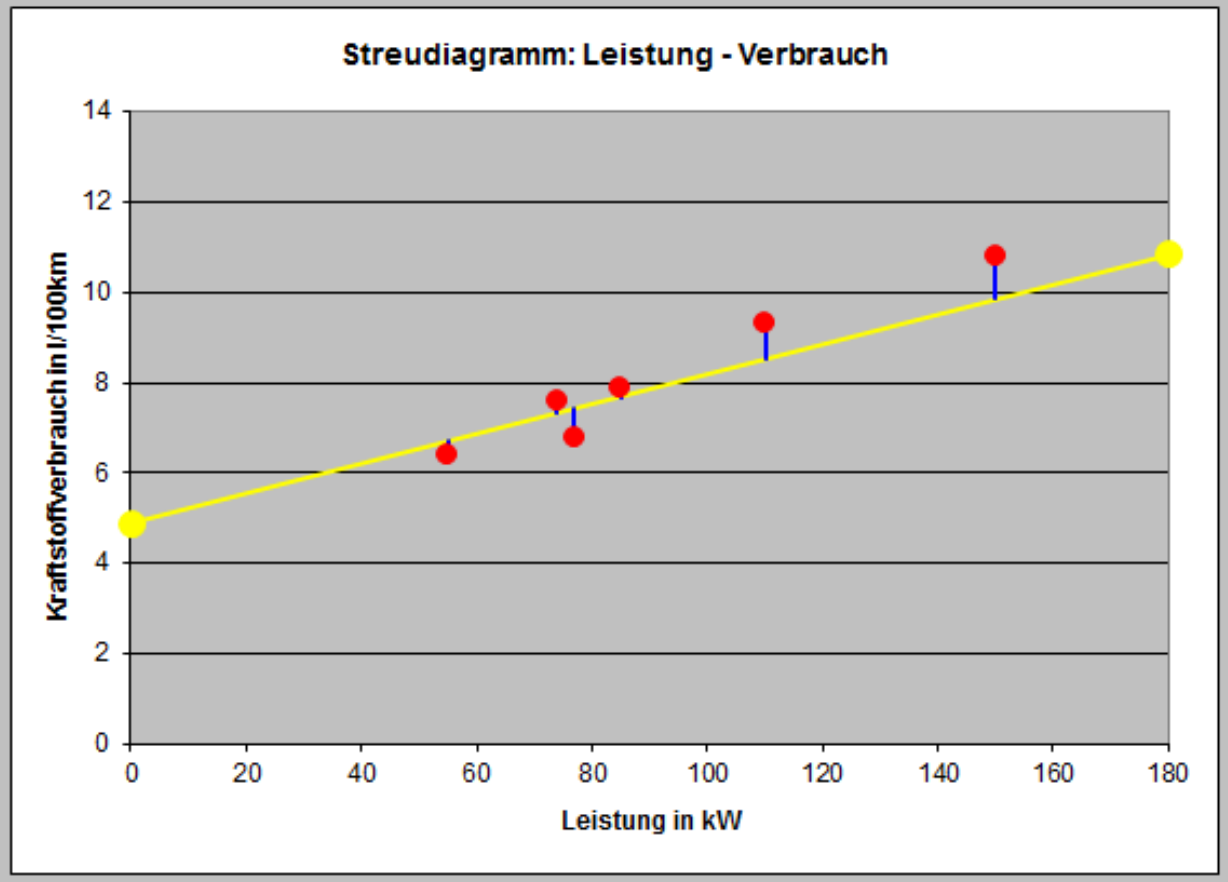
$$S = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2 \rightarrow \min!$$

Anpassung nach dem Prinzip der Kleinsten Quadrate

Die Aufgabe ist wieder, die gelbe Gerade mit Hilfe der beiden Schieberegler den roten Punkten möglichst gut anzupassen.

Möglichst gut bedeutet, dass die Summe der *quadratierten Abweichungen* e_i *(blaue Linien) von der Geraden minimiert* werden soll.

Die Summe der quadrierten Abweichungen wird zur Kontrolle im blauen Kasten angegeben.



Abstand: $b_0 = 4{,}87$

Steigung: $b_1 = 0{,}033$

Summe der $e_i^2 = 2{,}12$

Regressionsgerade einblenden

Regressionsgerade ausblenden

links rechts

Analytische Herleitung

$$\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) = 0 \qquad \frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)x_i = 0$$

$$\text{(i)} \qquad \sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\text{(ii)} \qquad \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^{\,2}$$

Aus (i) $\Rightarrow$ $\hat{b}_0 = \bar{y} - b_1 \bar{x}$

nach Substitution:

$$\hat{b}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Interpretation der Formeln für Koeffizienten

- Steigung der Regressionsgerade:

  Kovarianz von X und Y dividiert durch die Varianz von X

$$\hat{b}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

- Abstand auf der Ordinate:

  Lineare Regressionsgerade verläuft durch den Schwerpunkt der Punkte

$$\hat{b}_0 = \bar{y} - b_1 \bar{x}$$

Tabellarisches Rechenschema

Nr.	Xi	Yi	Xi^2	Xi*Yi	Yi^2
1	55	6,4	3025	352	40,96
2	74	7,6	5476	562,4	57,76
3	77	6,8	5929	523,6	46,24
4	85	7,9	7225	671,5	62,41
5	110	9,3	12100	1023	86,49
6	150	10,8	22500	1620	116,64
Summe	551	48,8	56255	4752,5	410,5

Mittelwert von X: 91,83
Mittelwert von Y: 8,13

$$\hat{b}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Berechnung von b_1:

Nenner 33929,00

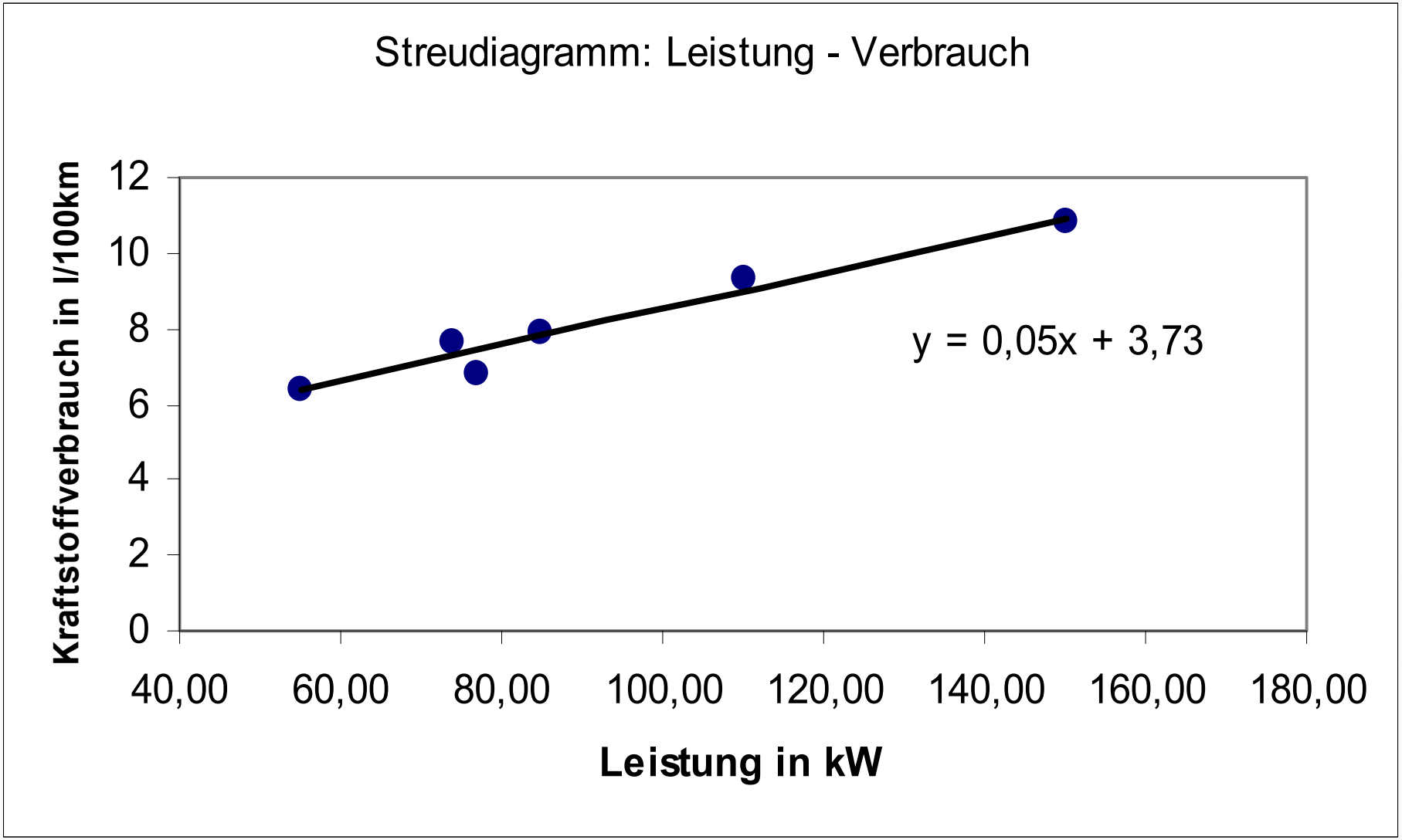
Zähler 1626,20

b_1 = 0,05

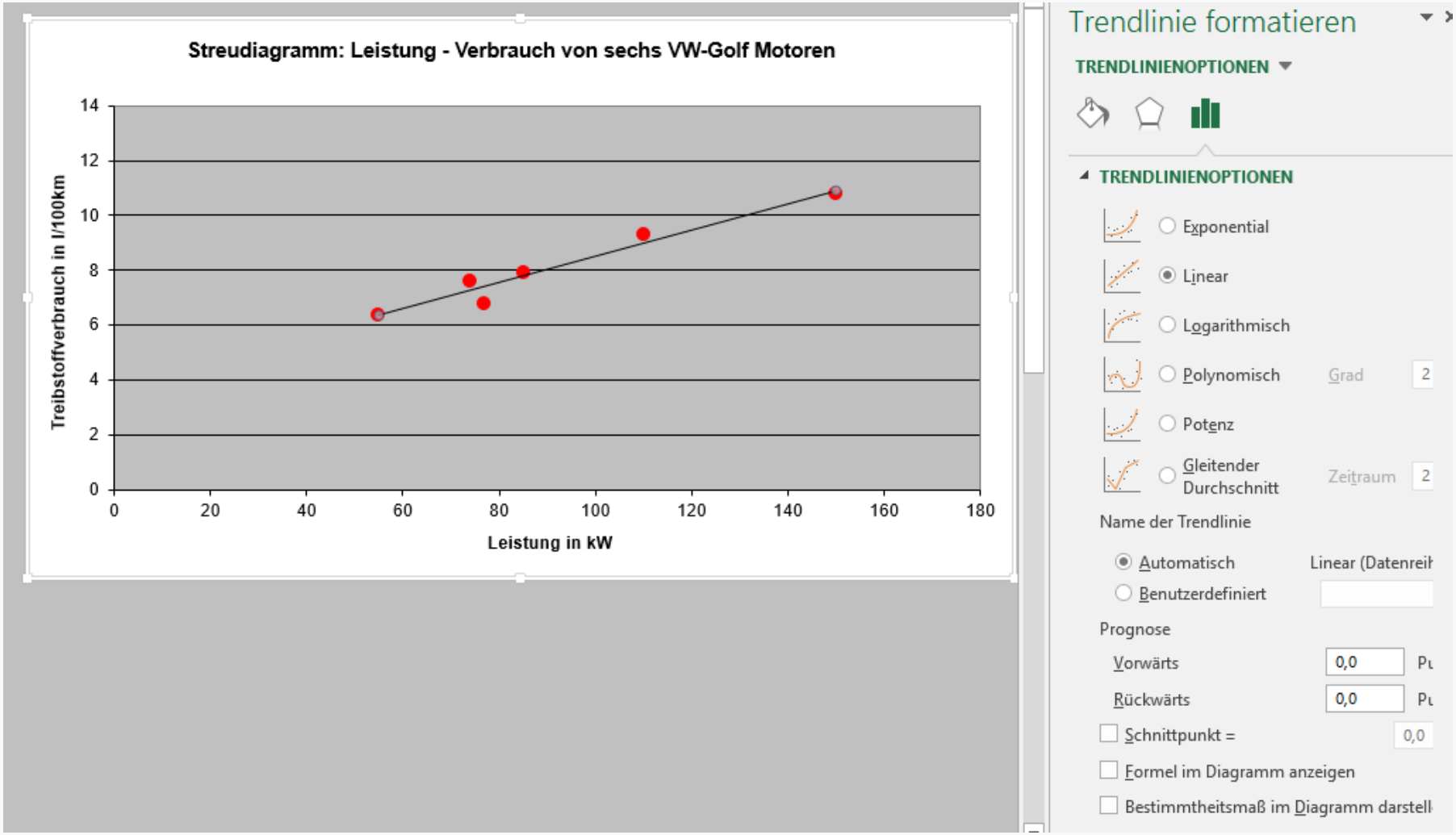
b_0 = 3,73

$$\hat{b}_0 = \bar{y} - b_1 \bar{x}$$

Graphische Darstellung



Automatisierte Berechnung mit EXCEL



Regressionsgerade als Instrument zur Vorhersage

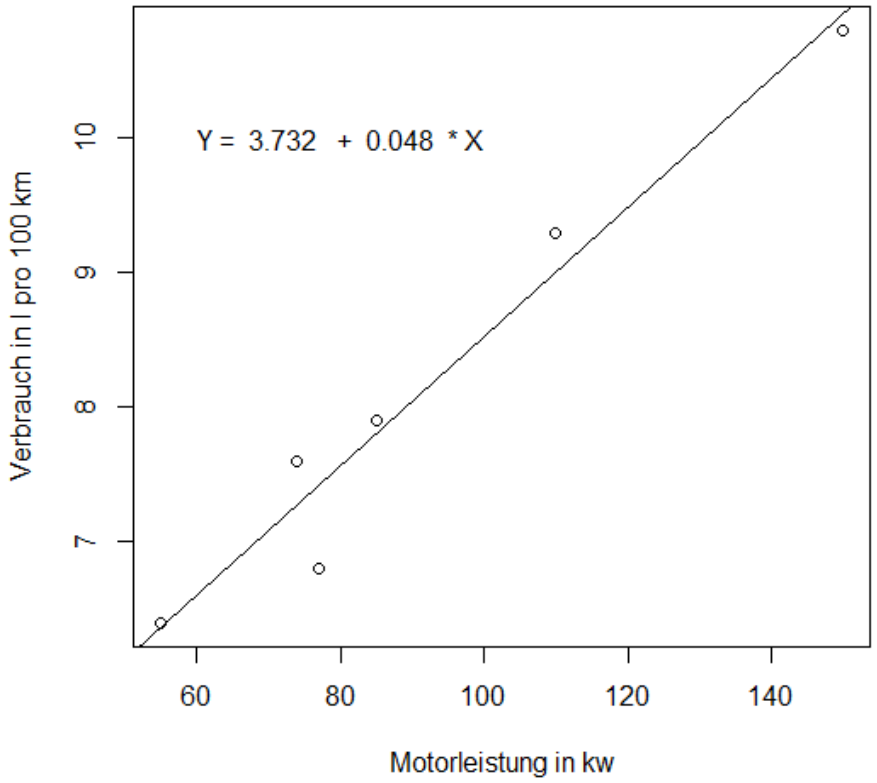
Basierend auf den geschätzten Parametern können wir für einen x Wert den zugehörigen y Wert schätzen

Prognose-Szenarien	
Wert von x	Schätzwert für y
40	5,65

Analyse mit R

```
> # Regression via lm
> # ========================================================
>
> res.lm <- lm(fuel ~ kw)
> res.lm$coeff                  # Zugriff auf Komponente c
(Intercept)          kw
  3.7318076   0.0479295
> coefficients(res.lm)          # Zugriff mit generischer
(Intercept)          kw
  3.7318076   0.0479295
> plot(kw, fuel, xlab="Motorleistung in kw",
+                          ylab="fuel in l pro 100 km")
> abline(res.lm)
> text(80, 10, paste("Y = ", round(res.lm$coeff[1], 3),
+             "  + ", round(res.lm$coeff[2], 3), " * X"))
>
> summary(res.lm)

Call:
lm(formula = fuel ~ kw)

Residuals:
       1        2        3        4        5        6
 0.03207  0.32141 -0.62238  0.09418  0.29595 -0.12123

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.731808   0.499881   7.465 0.001721 **
kw          0.047929   0.005163   9.284 0.000749 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3882 on 4 degrees of freedom
Multiple R-squared:  0.9557,    Adjusted R-squared:  0.9446
F-statistic:  86.2 on 1 and 4 DF,  p-value: 0.0007487
```

```
> anova(res.lm)
Analysis of Variance Table

Response: fuel
          Df  Sum Sq Mean Sq F value    Pr(>F)
kw         1 12.9905 12.9905  86.195 0.0007487 ***
Residuals  4  0.6028  0.1507
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Wichtige Eigenschaften der Regressionsgeraden

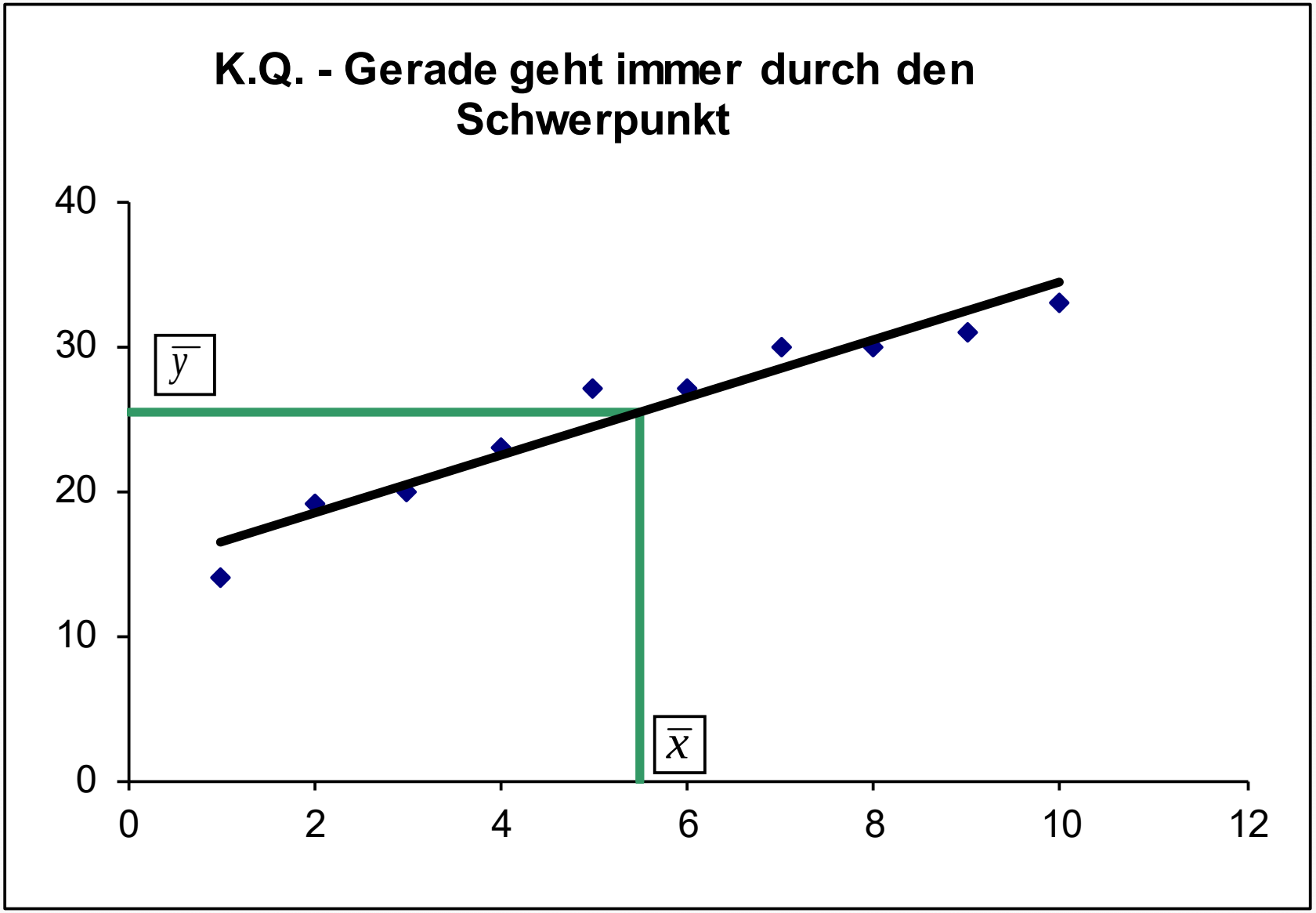
- „Fehlerausgleichende Gerade“

$$\sum_{i=1}^{n} e_i = 0$$

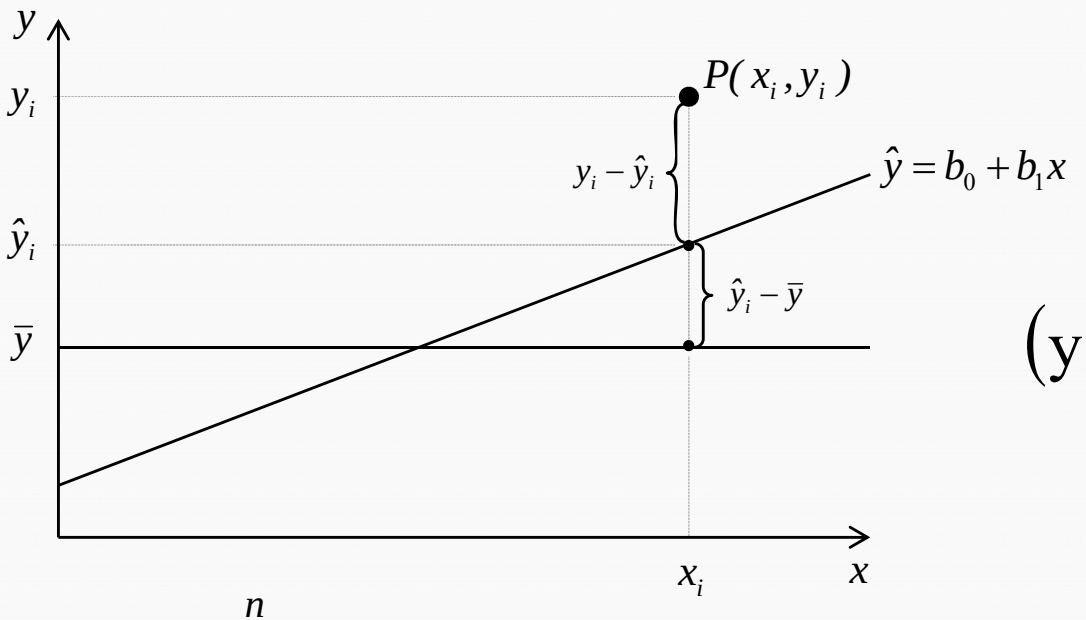
Die Summe der Abweichungen von der nach dem KI. Quadrate Prinzip optimalen Geraden ist gleich Null.

- Regressionsgerade läuft immer durch den Schwerpunkt der Punktwolke

$$\bar{y} = b_0 + b_1 \bar{x}$$



Variabilität der Regression



$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad \forall i$$

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

Totale Quadratsumme der Abweichungen vom arithmetischen Mittel

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

nicht erklärte (residuale) Abweichungsquadratsumme

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

erklärte Abweichungsquadratsumme

Zerlegung der Quadratsummen

$$SQT = SQR + SQE$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

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

r^2 = Bestimmtheitsmaß

Anteil der erklärten Varianz an der gesamten Varianz

r = Korrelationskoeffizient

Interpretation von r^2

- r^2 kann Werte zwischen
  - Null (kein Zusammenhang zwischen Y und X)

  und
  - Eins (alle Punkte liegen exakt auf einer Geraden)

  annehmen
- Je näher r^2 bei eins liegt, desto besser wird Y durch X mittels einer linearen Regression erklärt
- r^2 ist der Anteil der Variation von Y, der durch X erklärt werden kann

Bestimmung von r^2 im Beispiel

Nr.	Xi	Yi	Xi^2	Xi*Yi	Yi^2	$\hat{y}_i$	e_i	e_i^2	$(y_i-\bar{y})^2$	$(\hat{y}_i-\bar{y})$	$(\hat{y}_i-\bar{y})^2$
1	55	6,4	3025	352	40,96	6,37	0,03	0,00	3,00	-1,77	3,12
2	74	7,6	5476	562,4	57,76	7,28	0,32	0,10	0,28	-0,85	0,73
3	77	6,8	5929	523,6	46,24	7,42	-0,62	0,39	1,78	-0,71	0,51
4	85	7,9	7225	671,5	62,41	7,81	0,09	0,01	0,05	-0,33	0,11
5	110	9,3	12100	1023	86,49	9,00	0,30	0,09	1,36	0,87	0,76
6	150	10,8	22500	1620	116,64	10,92	-0,12	0,01	7,11	2,79	7,77
Summe	551	48,8	56255	4752,5	410,5	48,80	0,00	0,60	13,59	0,00	12,99

Mittelwert von X: 91,83
Mittelwert von Y: 8,13

Berechnung von b1:

Nenner 33929,00 $\quad b_1 = \mathbf{0,05}$

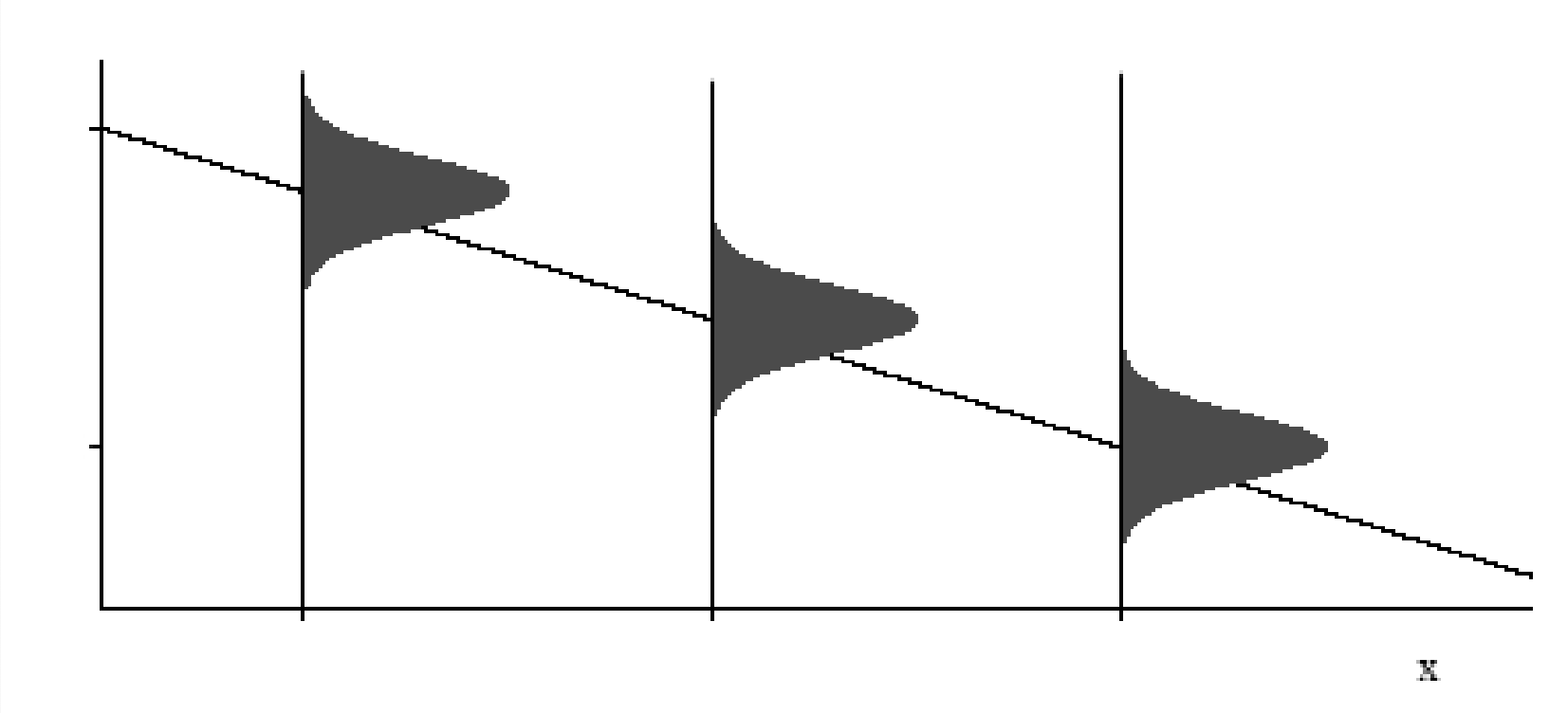
Zähler 1626,20 $\quad b_0 = \mathbf{3,73}$

SQT	13,59	100,0%
SQR	0,60	4,4%
SQE	12,99	95,6%

Inferenzstatistik

- Es wird angenommen, daß die Werte der unabhängigen Variablen feste (nichtzufällige) Größen sind.
- Es wird angenommen, daß sich die Beobachtungen der abhängigen Variablen durch einen in X linearen Term plus einer zufälligen Störkomponente ergeben.
- Über die Störkomponente werden folgende Annahmen getroffen
  - Keine systematische Störung, d.h. Erwartungswert ist null $E(e_i) = 0$
  - Konstante Streuung der Störkomponente $Var(e_i) = const.$
  - Die Störungen sind unabhängig voneinander $Cov(e_i, e_j) = 0$
  - Die Störkomponente sei normalverteilt mit Erwartungswert 0 und der Varianz σ^2

Modellannahme



Die bedingten Dichten von Y für gegebenen Wert von X unterscheiden sich nur in ihrem Erwartungswert

Modell der linearen Einfachregression

$f_{Y|x}$ ist Normalverteilungsdichte $N(\mu_x, \sigma^2)$.

σ^2 hängt nicht von x ab; (μ_x in der Regel schon)

Bedingter Erwartungswert ist lineare Funktion:

$$h_Y(x) = \mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x,$$

Koeffizienten β_0 und β_1 unbekannt.

Koeffizienten β_0 und β_1 werden aus unabhängigen Beobachtungspaaren $(X_i,\ Y_i)$ $1 \leq i \leq n$ geschätzt.

Maximum Likelihood Schätzung

Likelihood Funktion:

$$L = \prod_{i=1}^{n} f_{Y|x_i}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2}$$

- Log-Likelihood:

Es wird deutlich, dass die Schätzung nach dem Maximum-Likelihood Prinzip zu exakt den selben Schätzergebnissen, wie das kleinste Quadrate Prinzip führt!

$$-2\ln L = n \cdot \ln(2\pi) + n \cdot \ln(\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Eigenschaften der Maximum Likelihood-Schätzer

$\hat{\beta}_0$ und $\hat{\beta}_1$ beide normalverteilt

$\hat{\beta}_0$ und $\hat{\beta}_1$ beide erwartungstreue Schätzer:

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ ist ein erwartungstreuer Schätzer der *Fehlervarianz* σ^2.

Signifikanz der Regressionsbeziehung

- Frage ist der Anteil der erklärten Varianz signifikant?
- Antwort: F-Test
- Erklärte durch nichterklärte mittlere Quadratsumme (das ist die Quadratsumme durch die Zahl der Freiheitsgrade dividiert)
- Diese Prüfgröße ist bei Gültigkeit der Nullhypothese, dass kein Erklärungswert vorliegt, F-verteilt mit *1* und *n-2* Freiheitsgraden

$$F = \frac{SQE/1}{SQR/(n-2)} = \frac{r^2/1}{(1-r^2)/(n-2)}$$

Durchführung des Tests

ANOVA	(Analysis of Variance)				
	Freiheitsgrade (df)	*Quadratsummen (SS bzw. SQ)*	*Mittlere Quadratsumme*	*Prüfgröße (F)*	*P-Wert*
Regression	1	12,990	12,990	86,195	0,0007
Residuen	4	0,603	0,151		
Gesamt	5	13,593			
	r^2=	0,956			
	$(1-r^2)$=	0,044			
	$(1-r^2)/4$=	0,011			

Schätzung von σ^2

Die Schätzung der unbekannten Varianz der Störkomponente ist die Voraussetzung für Inferenz über die Parameter bzw. für Konfidenzintervalle für Prognosewerte.

Naheliegend ist die nachstehende Formel (E(e)=0!)

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

Fehlervarianz

Die positive Quadratwurzel führt zum Standardfehler der Residuen (Residual Standard Error)

Konfidenzintervalle und Tests für Regressionskoeffizienten

Schätzung der Varianz der Regressionskoeffizienten

$$\hat{\sigma}_{b_0}^{\ 2} = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\hat{\sigma}^2$$

$$\hat{\sigma}_{b_1}^{\ 2} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Interpretation:

Bei großen Werten von x ist die Variabilität der Konstanten ceteris paribus größer.

Je stärker die x-Werte streuen, desto geringer ist ceteris paribus die Streuung beider Koeffizienten

Konfidenzintervalle und Tests für Regressionskoeffizienten

$$P(b_i - \hat{\sigma}_{b_i} t_{n-2;1-\alpha/2} \leq \beta_i \leq b_i + \hat{\sigma}_{b_i} t_{n-2;1-\alpha/2}) = 1-\alpha \quad i = 0,1$$

1-α Konfidenzintervall für die Regressionskoeffizienten:

$$[b_i - \hat{\sigma}_{b_i} t_{n-2;1-\alpha/2} \leq \beta_i \leq b_i + \hat{\sigma}_{b_i} t_{n-2;1-\alpha/2}]$$

H_0	H_1	H_0 wird verworfen, falls
$\beta_i = \beta_i^0$	$\beta_i \neq \beta_i^0$	$\left\|(b_i - \beta_i^0) / \hat{\sigma}_{b_i}\right\| > t_{n-2;1-\alpha/2}$
$\beta_i \leq \beta_i^0$	$\beta_i > \beta_i^0$	$(b_i - \beta_i^0) / \hat{\sigma}_{b_i} > t_{n-2;1-\alpha}$
$\beta_i \geq \beta_i^0$	$\beta_i < \beta_i^0$	$(b_i - \beta_i^0) / \hat{\sigma}_{b_i} < t_{n-2;1-\alpha}$

Der Test für die Steigung ist äquivalent zum Test auf Korrelation!

Test für die Regressionskoeffizienten

- Nullhypothese: $b_i=0$

	Koeffizient	*Standardfehler*	*t-Statistik*	*P-Wert*
b_0	3,732	0,500	7,465	0,0017
b_1	0,048	0,005	9,284	0,0007

$$\hat{\sigma}^2 = 0,60 / 4 = 0,15$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 56.255 - 6\,91,83^2 = 5.654,8$$

$$\hat{\sigma}_{b_1}^2 = \sqrt{\frac{0,15}{5.654,8}} = 0,005$$

- Interpretation:
  - $b_0=0$ ... Geht die Regression durch den Ursprung?
  - b1=0 ... Ist die Steigung signifikant von Null verschieden?

    Das entspricht im Fall der Einfachregression der zuvor diskutierten Fragestellung:

    Ist der Anteil der erklärten Varianz signifikant?

    Hinweis: Vergleiche den p-value für die Steigung mit dem F-Test

Überblick über Testverfahren

- Test auf Korrelation zwischen x und y

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{mit} \quad n-2 \quad \text{Freiheitsgraden}$$

- F-Test auf Erklärungswert von x in Bezug auf y

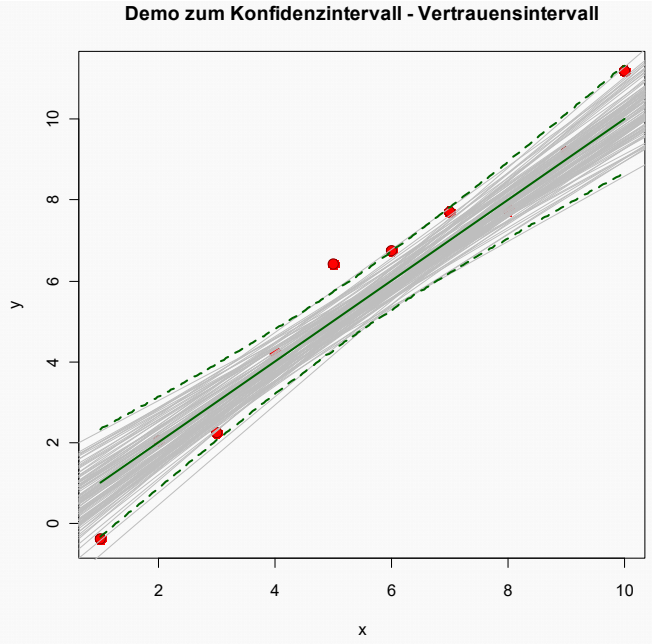
$$F = \frac{SQE/1}{SQR/(n-2)} = \frac{r^2/1}{(1-r^2)/(n-2)}$$ bei Gültigkeit der Nullhypothese F-verteilt mit *1* und *n-2* Freiheitsgraden

- Test auf Signifikanz der Steigung
  $H_0: b_1=0$ $H_A: b_1\neq 0$ Teststatistik $b_1 / \hat{\sigma}_{b_1}$ ist unter der Nullhypothese t-verteilt mit n-2 Freiheitsgraden
- Im Fall der linearen Einfach-Regression sind diese 3 Tests äquivalent

Wie genau sind Prognosen?

- *Prognose für den Mittelwert:* Welchen Wert nehmen Beobachtungen der abhängigen Variable bei einem bestimmten Wert der unabhängigen Variablen im Mittel an?

- *Individuelle Prognose:* Welchen Wert nimmt eine neue individuelle Beobachtung an einem Punkt *x* an?



Konfidenzintervall für den durchschnittlichen Prognosewert

$$T = \frac{\hat{Y}_i - E(Y_i)}{S_{\hat{Y}_i}}$$

t-verteilt mit n-2 Freiheitsgraden

$$\rightarrow \; P(\hat{Y}_i - tS_{\hat{Y}_i} \leq E(Y_i) \leq \hat{Y}_i + tS_{\hat{Y}_i}) = 1 - \alpha$$

Für eine konkrete Stichprobe ergibt sich damit das folgende **Konfidenzintervall für den durchschnittlichen Prognosewert (Vertrauensintervall)**

$$\hat{y}_i - ts_{\hat{Y}_i} \leq E(Y_i) \leq \hat{y}_i + ts_{\hat{Y}_i}$$

mit $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ *und* $s_{\hat{Y}_i} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}}$

Prognoseintervall für individuellen Prognosewert Y_i

$$T = \frac{\hat{Y}_i - Y_i}{S_F}$$

t-verteilt mit n-2 Freiheitsgraden

$$\rightarrow \; P(\hat{Y}_i - tS_F \leq Y_i \leq \hat{Y}_i + tS_F) = 1 - \alpha$$

Aus einer konkreten Stichprobe ergibt sich somit das folgende **Konfidenzintervall für die Prognose eines bestimmten Einzelwertes an der Stelle x_i:**

$$\hat{y}_i - ts_F \leq Y_i \leq \hat{y}_i + ts_F$$

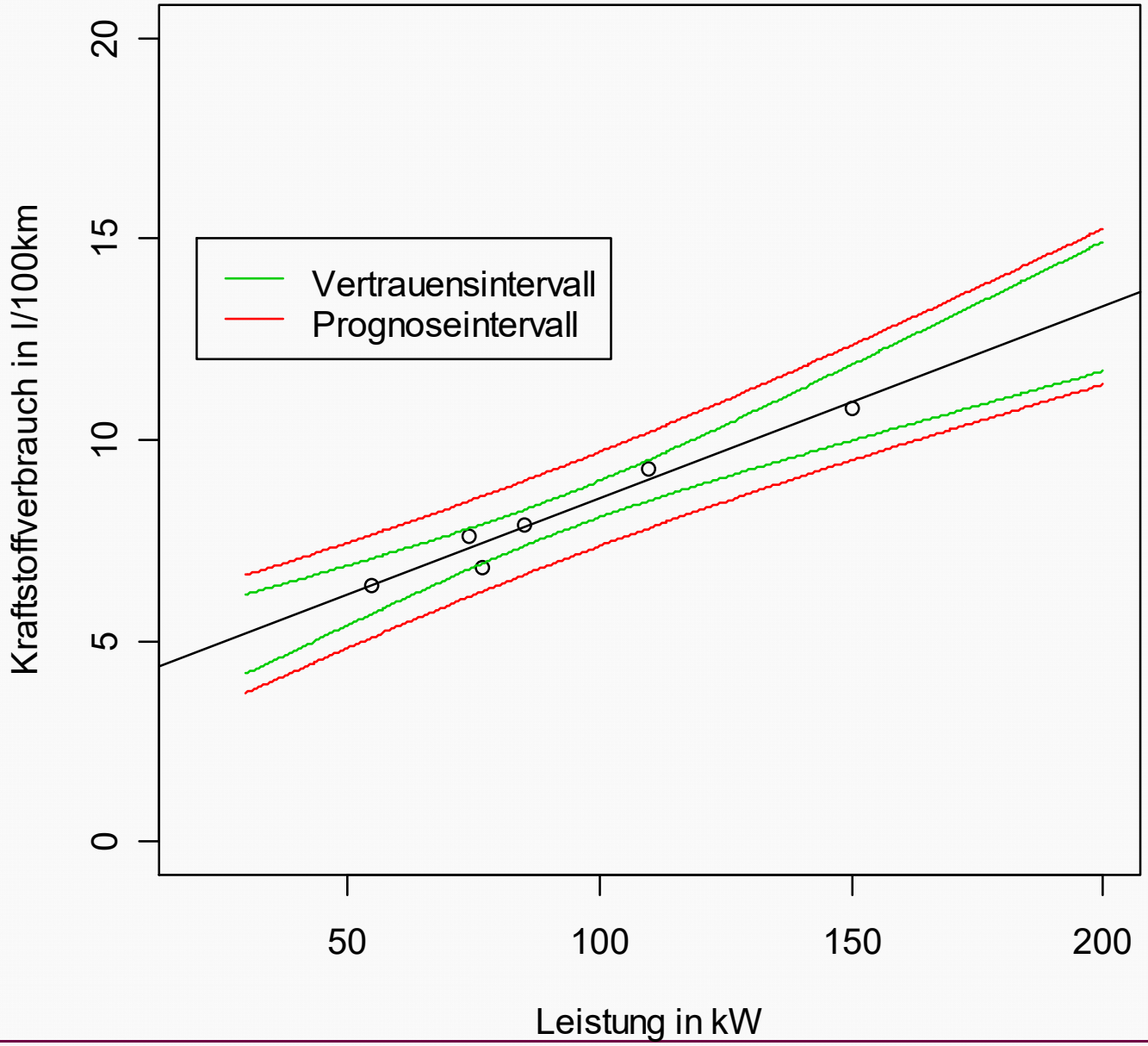
mit $\hat{y}_i = b_0 + b_1 x_i$ *und* $s_F = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}$

Anwendung im Beispiel

Prognoseintervall für Einzelwerte (individuelle Prognosewerte)

		xi	S_F	Prognose	UG	OG
se	0,388	50	0,4717	6,12828	4,818722	7,437843
t:	2,78	55	0,4604	6,36793	5,089596	7,646264
		60	0,4504	6,60758	5,357137	7,858018
		65	0,4416	6,84723	5,621119	8,073331
		70	0,4342	7,08687	5,881326	8,292419
		75	0,4282	7,32652	6,137561	8,515479
		80	0,4237	7,56617	6,389658	8,742677
		85	0,4208	7,80582	6,637483	8,974147
		90	0,4194	8,04546	6,880947	9,209979
		95	0,4196	8,28511	7,120006	9,450214
		100	0,4214	8,52476	7,354668	9,694847
		105	0,4248	8,76441	7,584988	9,943822
		110	0,4297	9,00405	7,811069	10,19704
		115	0,4360	9,2437	8,033051	10,45435
		120	0,4438	9,48335	8,251113	10,71558

Vertrauens- und Prognoseintervall



Prognosegenauigkeit

- Genauigkeit der Prognosen an einer Stelle x_0 nimmt mit zunehmenden Abstand zwischen x_0 und dem Mittelwert $\bar{x}$ ab.
- Die Länge von Konfidenzintervallen nimmt mit zunehmenden Abstand zwischen x_0 und Mittelwert $\bar{x}$ zu.
- Vorsicht bei Prognosen bzw. Konfidenzintervallen außerhalb des Bereichs, in dem Daten vorliegen!

„Hidden extrapolation“ – „distance to model“

Regression durch den Ursprung

$(y_1, x_1) \ldots (y_n, x_n)$

$$y_i = c \cdot x_i + e_i$$

$$S = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - c \cdot x_i)^2 \Rightarrow \min!$$

$$\frac{\partial S}{\partial c} = -2 \sum_{i=1}^{n} (y_i - c \cdot x_i) x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i = c \sum_{i=1}^{n} x_i^{\,2}$$

$$c = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^{\,2}}$$

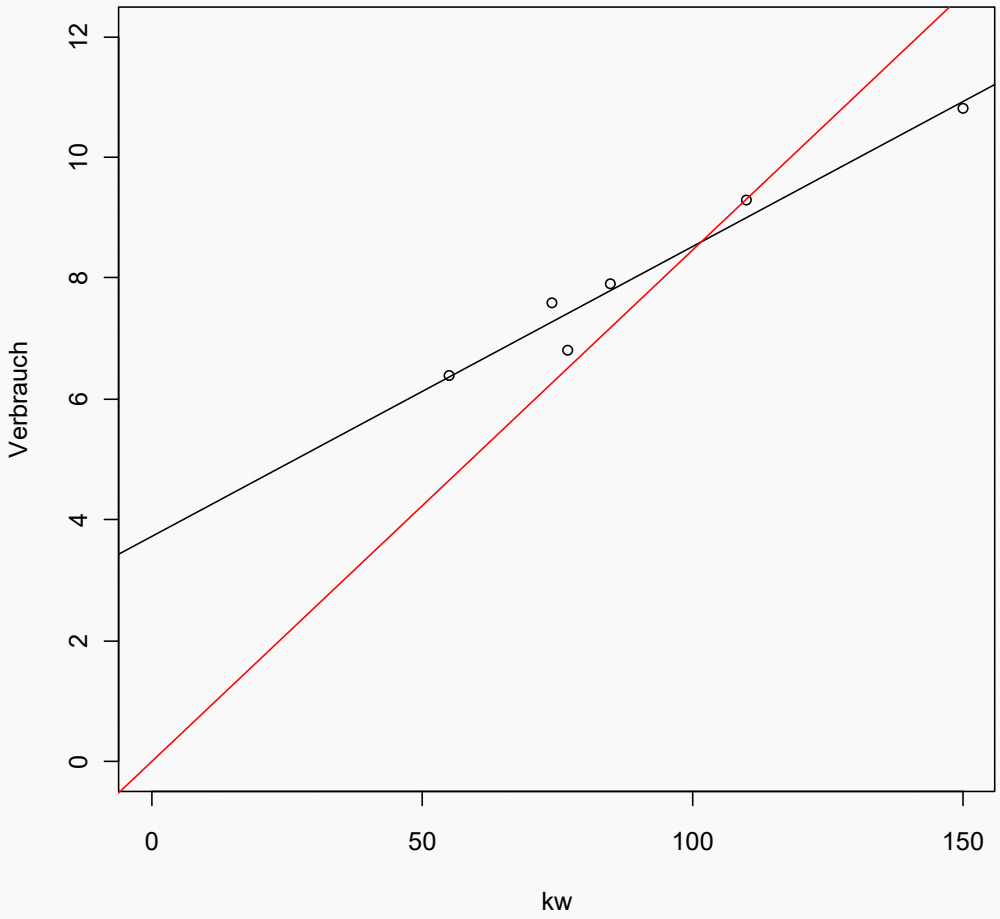
Spezialfall:

$$x_1 = \ldots = x_n = 1$$

$$c = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Das arithmetische Mittel ist ein Kleinster Quadrate Schätzer

Regression durch den Ursprung



Die schwarze Gerade ergibt sich aus dem Modell $y=b_0+b_1x$, die rote Gerade aus dem Modell $y=b_1x$

Im Zweifelfall empfiehlt es sich die Konstante auch wenn sie nicht signifikant ist im Modell zu belassen.

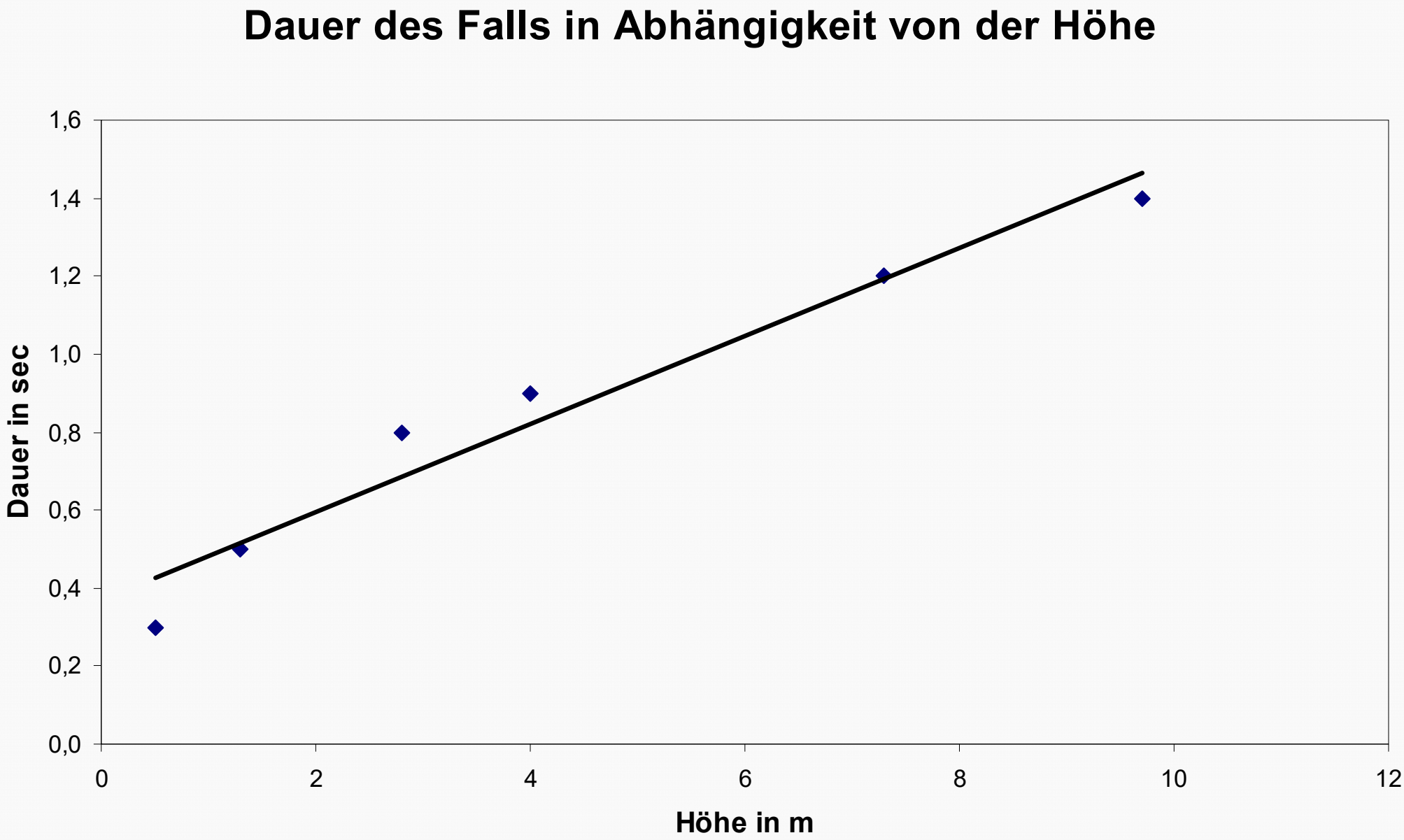
Nichtlineare Zusammenhänge

- Beispiel: Stoppen der Fallzeit eines Gegenstandes aus verschiedenen Höhen zur Bestimmung der Erdbeschleunigung
- Ein Gegenstand wird aus einer bestimmten Höhe fallen gelassen
- die Fallzeit in Sekunden wird gemessen
- Frage: Welcher Zusammenhang besteht zwischen der Fallzeit und der Höhe?

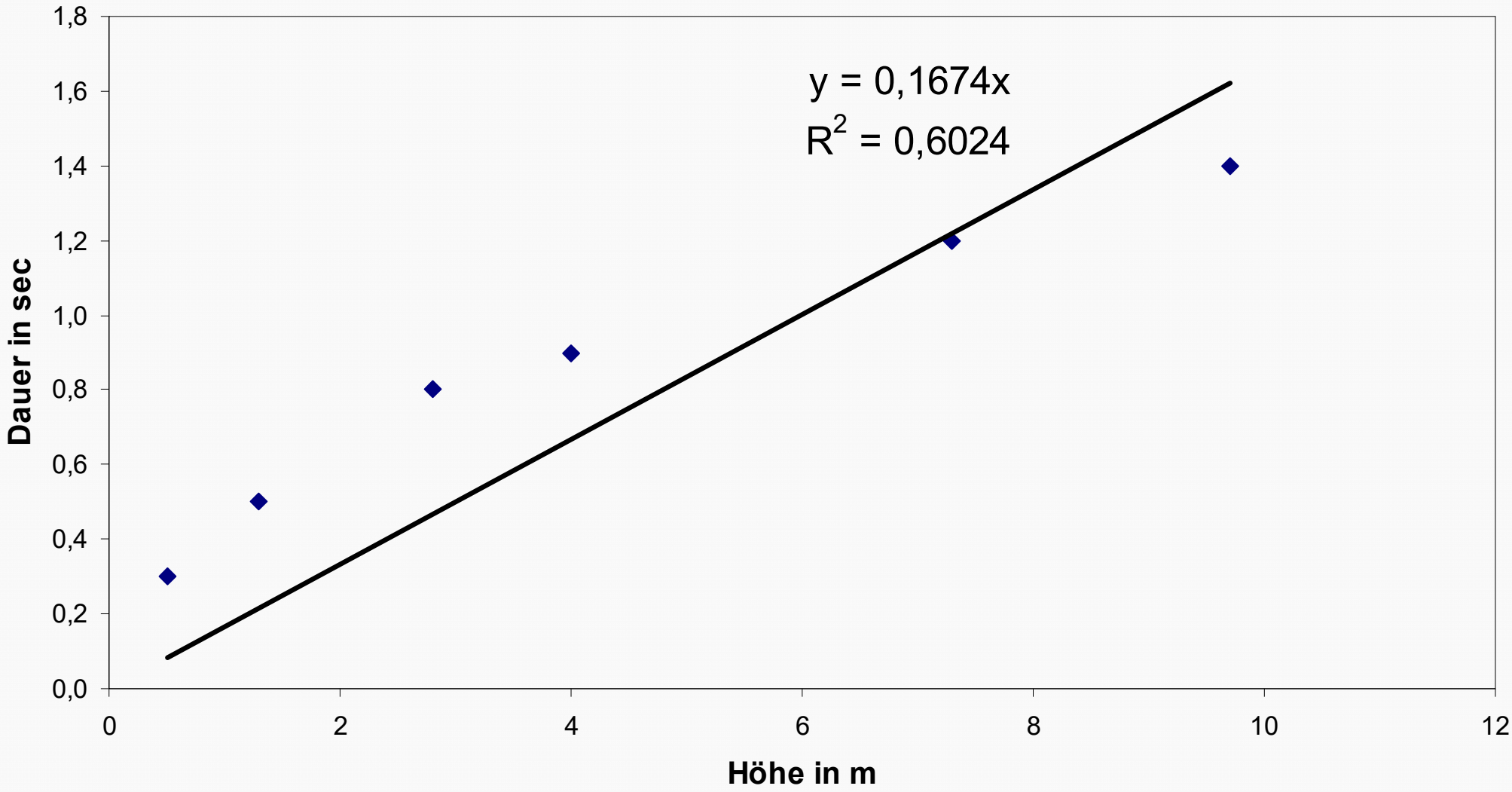
Meßergebnisse

Höhe	Fallzeit
0,50	0,30
1,30	0,50
2,80	0,80
4,00	0,90
7,30	1,20
9,70	1,40

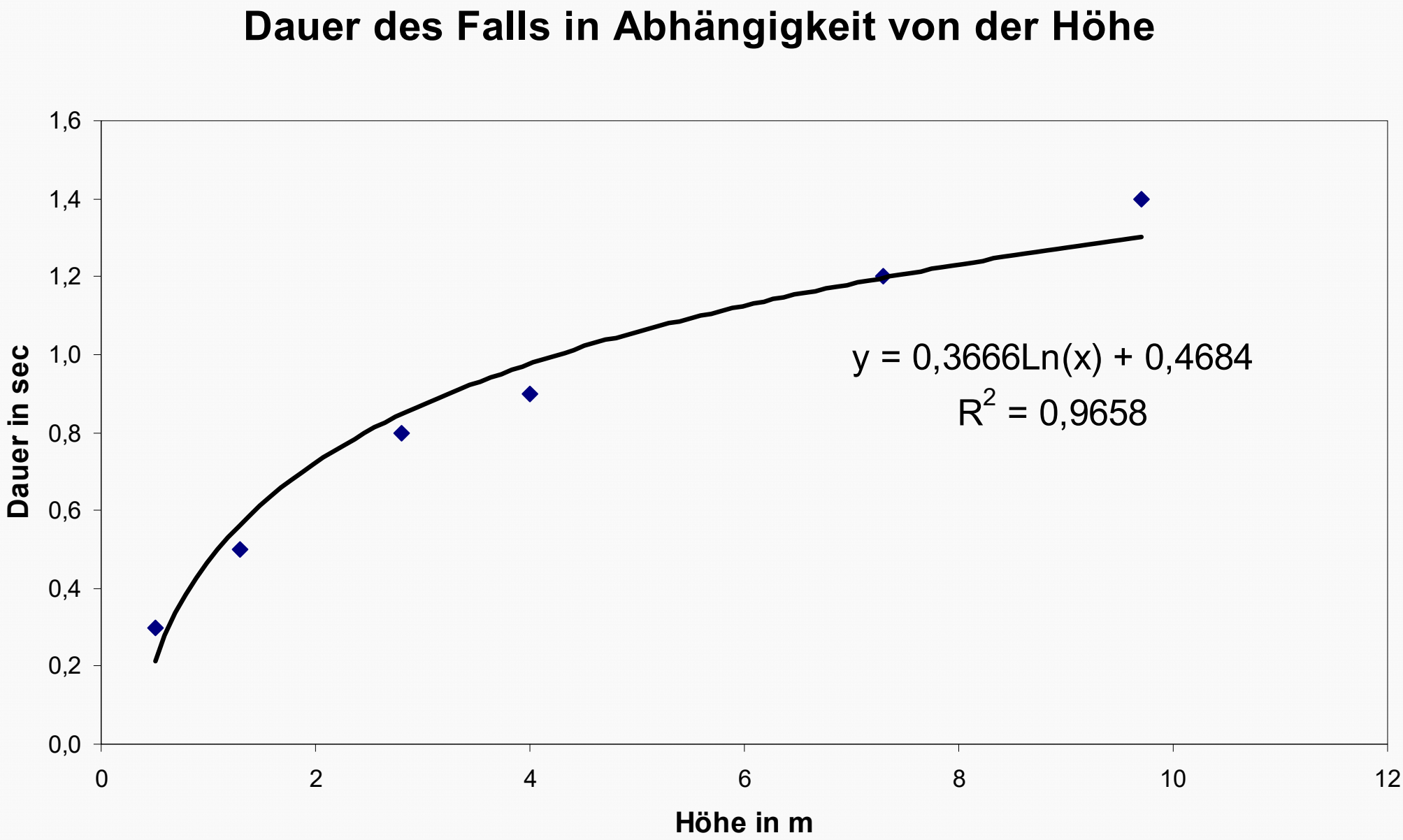
Lineare Trendlinie



Lineare Trendlinie durch den Nullpunkt



Logarithmische Trendlinie



Änderung der x-Achse

$h=g/2*t^2$ h...Höhe g...Erdkonstante t...Zeit

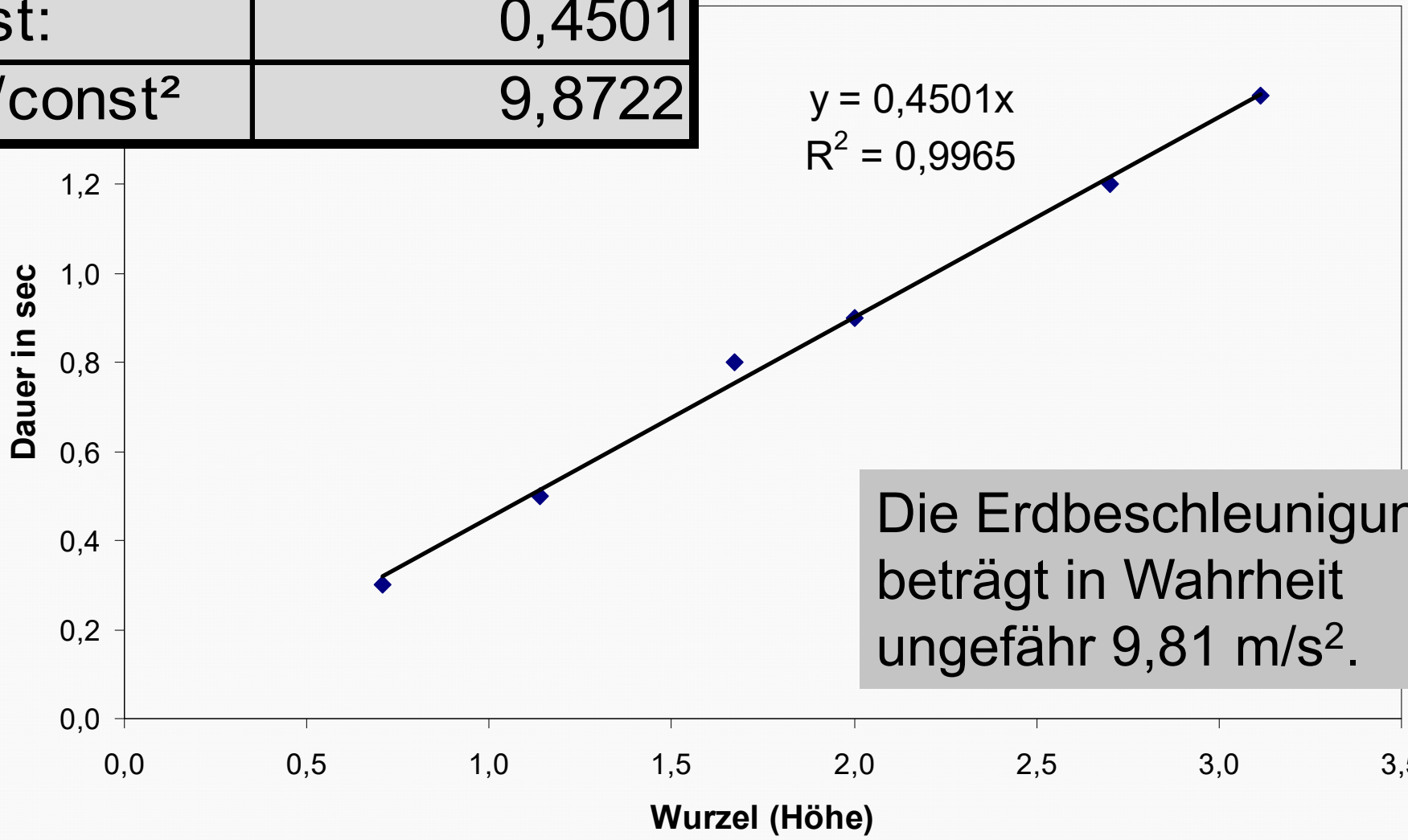
t=wurzel(2h/g) bzw. t=const.*wurzel(h)

const. = wurzel(2/g)

Wurzel(Höhe)	Fallzeit
0,71	0,30
1,14	0,50
1,67	0,80
2,00	0,90
2,70	1,20
3,11	1,40

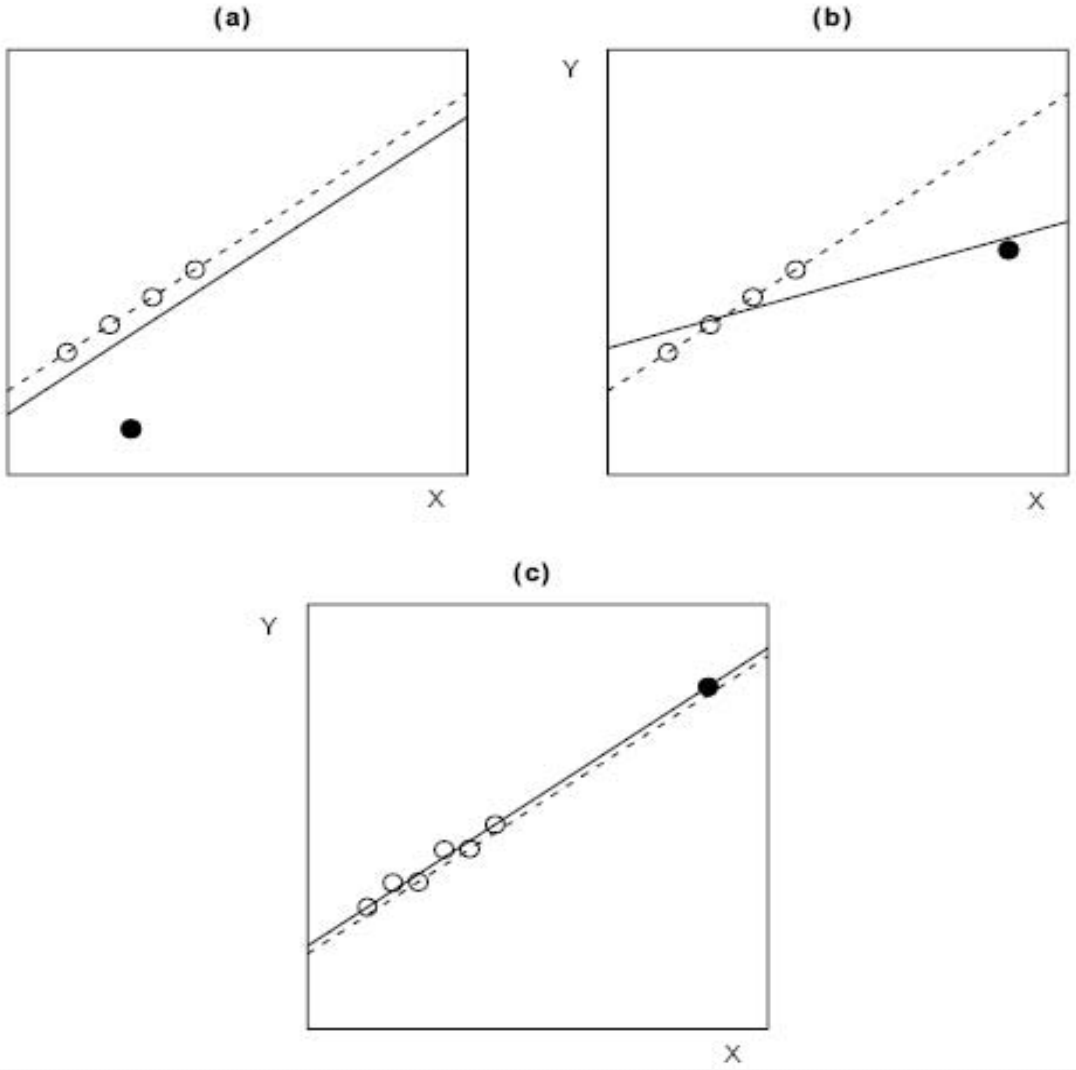
Änderung der x-Achse

const:	0,4501
g=2/const²	9,8722



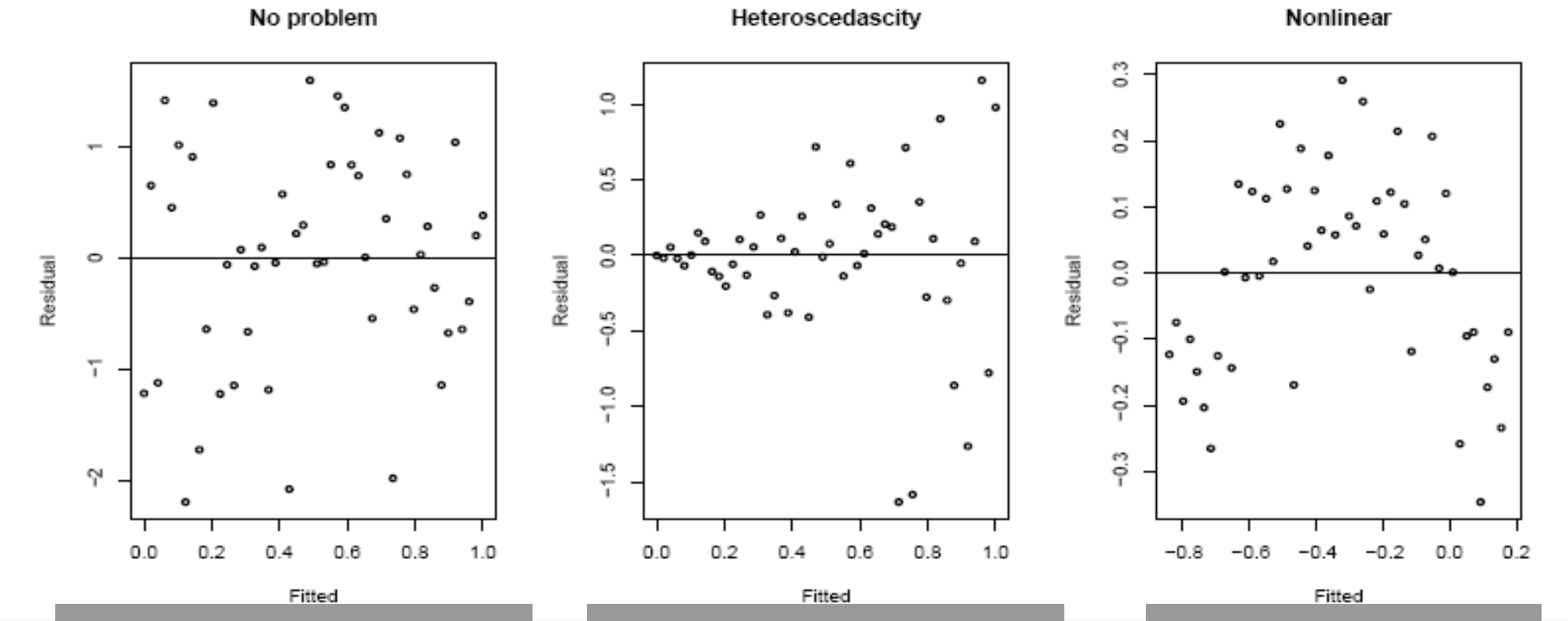
Die Erdbeschleunigung, beträgt in Wahrheit ungefähr 9,81 m/s².

Regression Diagnostics



- (a) Outlier not at a high leverage point and hence not influential.
- (b) Outlier at a high-leverage point and hence influential.
- (c) In-line at a high leverage point and hence not influential.
- Influence on coefficients = Leverage × Outlyingness

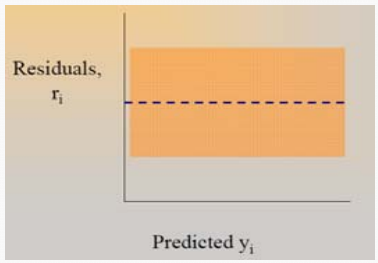
Residuen versus Fitted Values

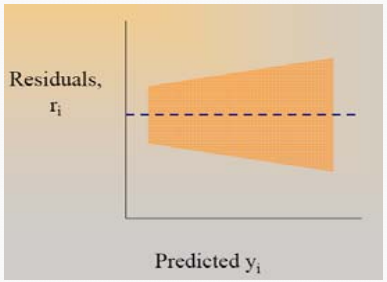


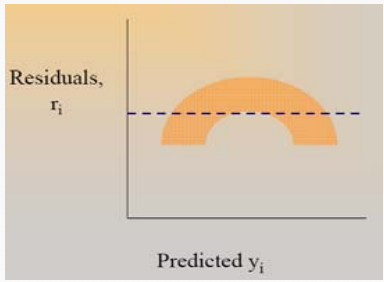
Ordinary Least Squares

Weighted Least Squares

Nonlinear Models







4 Datensätze von Anscombe

Datensatz 1		Datensatz 2		Datensatz 3		Datensatz 4	
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Die Lösungen der Regressionsgeraden sehen folgendermaßen aus:

a) $y = 3 + 0{,}5*x$

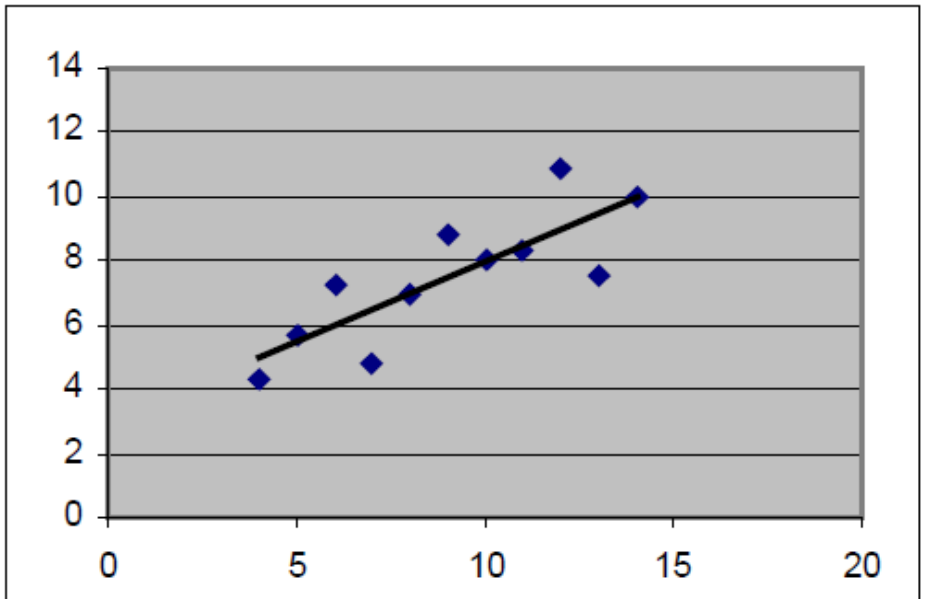
b) $y = 3 + 0{,}5*x$

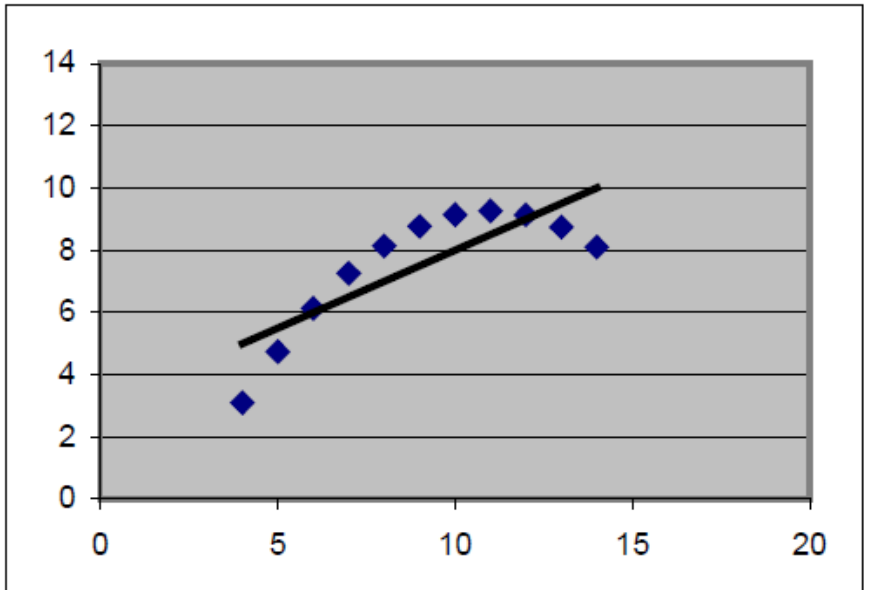
c) $y = 3 + 0{,}5*x$

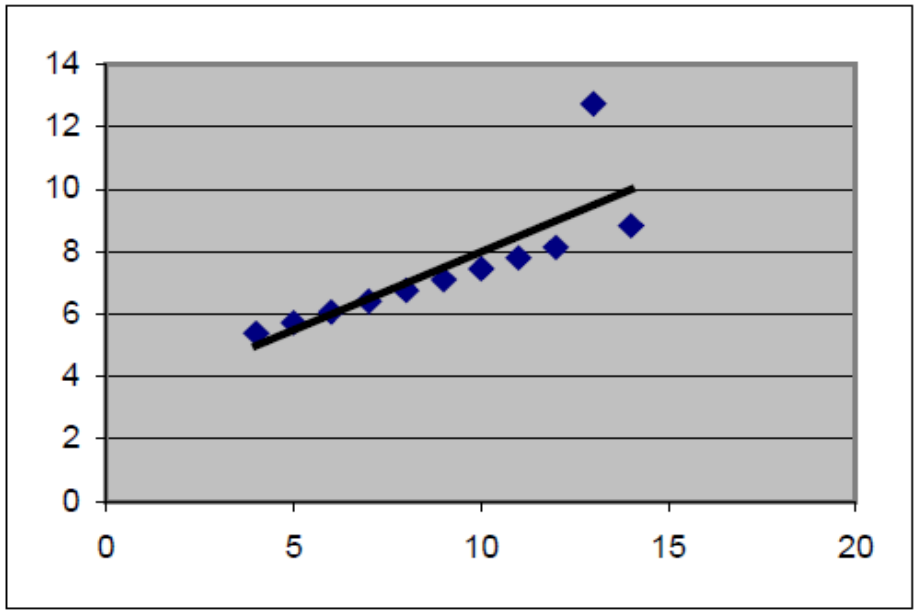
d) $y = 3 + 0{,}5*x$

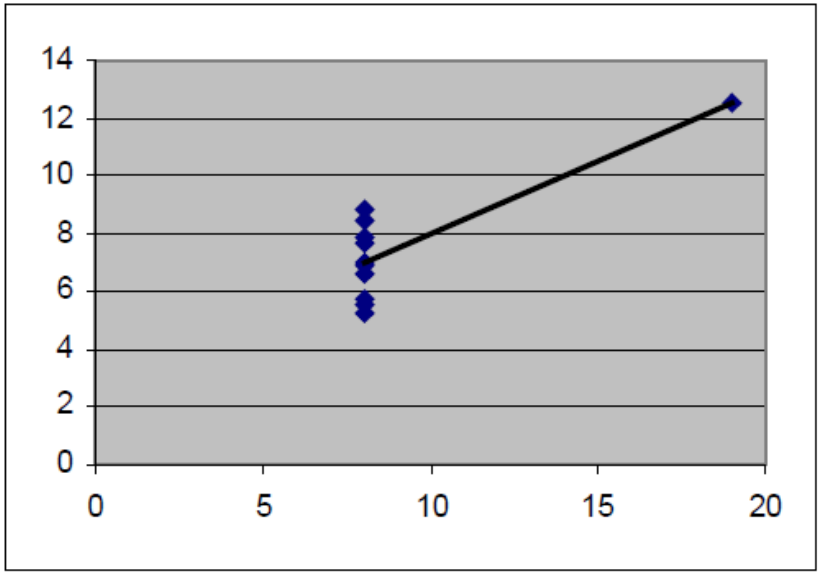
Für jeden Datensatz gilt, dass die Korrelation 0,816 und somit $R^2=0{,}667$ ist.

4 verschiedene Interpretationen





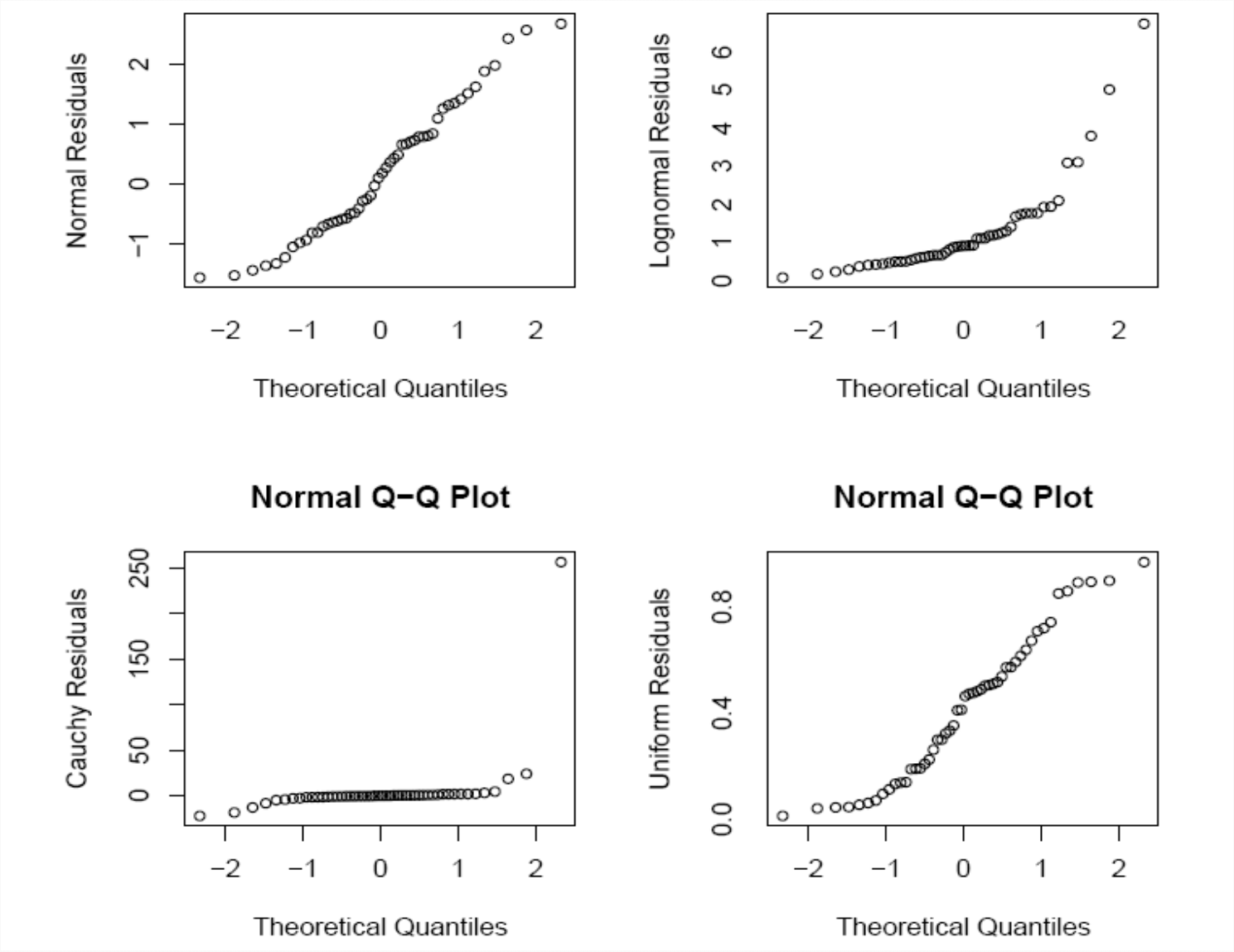




Nicht-Normalverteilte Fehler

- Bei nicht-normalverteilten Fehlern besteht die Möglichkeit sich bei großen Stichproben auf den zentralen Grenzwertsatz zu berufen.
- Allerdings verliert der Kleinst Quadrate-Schätzer an Effizienz, wenn Fehlerverteilungen vorliegen, bei denen große Abweichungen mit hoher Wahrscheinlichkeit vorkommen (heavy tailed distribution)
- Bei schiefen Verteilungen ist die Schätzung des bedingten Erwartungswertes E(Y|X) problematisch
- Multimodale Fehlerverteilungen deuten auf eine nichtberücksichtigte Heterogenität in der Population hin

QQ-Plots von Residuen für Simulierte Daten



Vorgehen bei Nicht-Normalverteilten Fehlern

- Für symmetrische Verteilungen, bei denen extreme Abweichungen nur mit sehr geringer Wahrscheinlichkeit auftreten, sind die Konsequenzen der Verletzung der Verteilungsannahmen häufig nicht allzu gravierend und können bei der praktischen Anwendung unberücksichtigt bleiben.
- Andernfalls kann man versuchen, durch geeignete Transformationen der Zielvariablen (log, Wurzelfunktion) eine bessere Anpassung an die Normalverteilung zu erzielen
- Nutzung alternativer Methoden:
  - Robust Regression Models
  - Generalized Linear Models