Statistik

TEIL 8

Hans-Hermann Thulke
ba @ thulke-statistics.de
0172-3449934

Statistik

- Daten erheben, verstehen, werten
- Hypothesen prüfen
- Modellieren von Zusammenhängen

Modellieren von Zusammenhängen

Multivariate Zufallsgrößen

Zusammenhangsmaße

Zusammenhangsmodelle

Trendbestimmung

Zusammenhang - modellieren

Unabhängige Merkmale
(Regressoren)

X_1

X_2

.
.
.

X_k

Abhängiges Merkmal
(Regressand)

Y

Mathematische Funktion
(Regressionsfunktion)

$\hat{Y} = f(X_1, X_2, \dots, X_k)$

lineare Einfachregression

$\hat{Y} = \beta_1 + \beta_2 X$

lineare Mehrfachregression

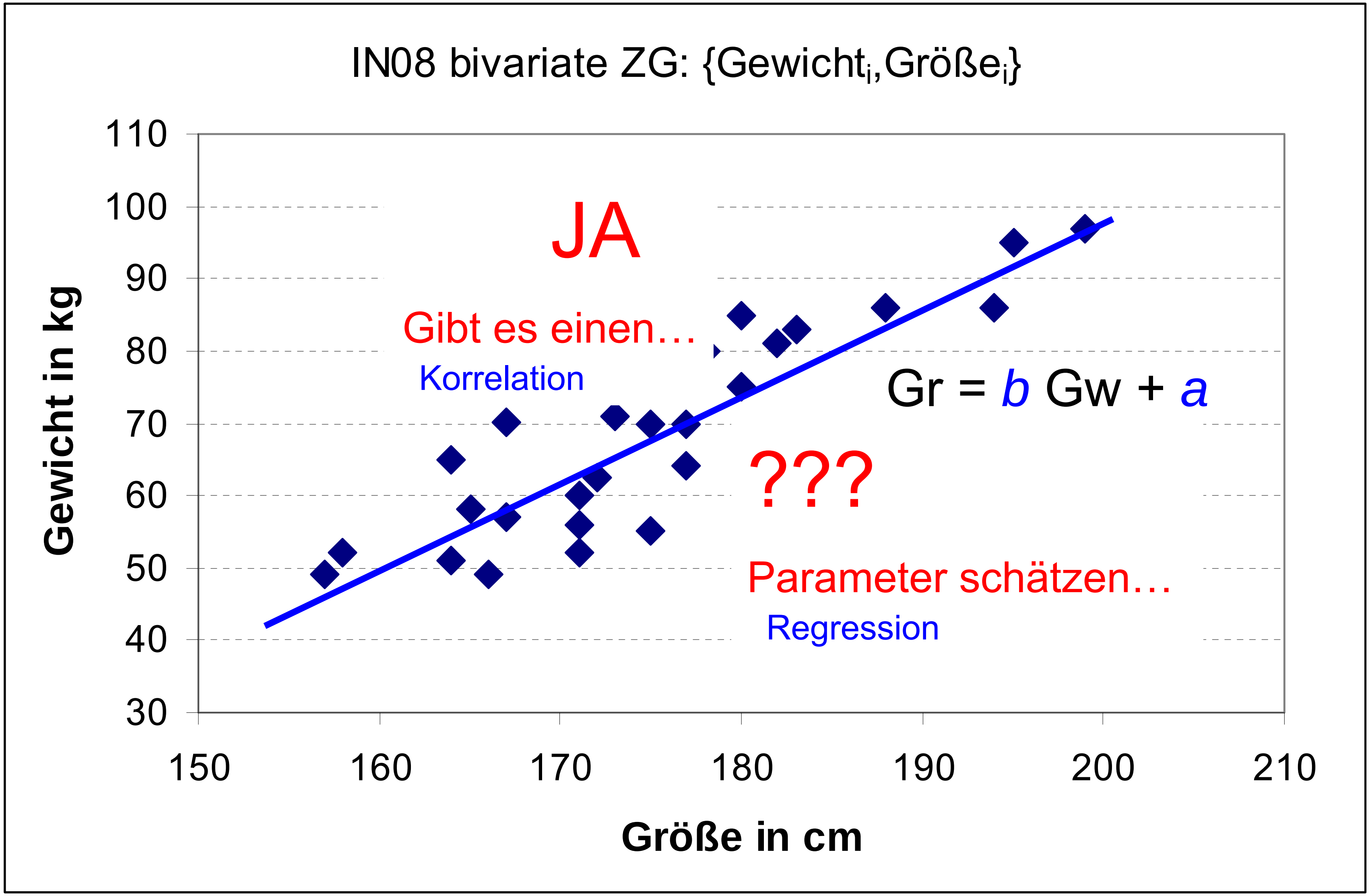
$\hat{Y} = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k$

nichtlineare Regression

Linearisierbar

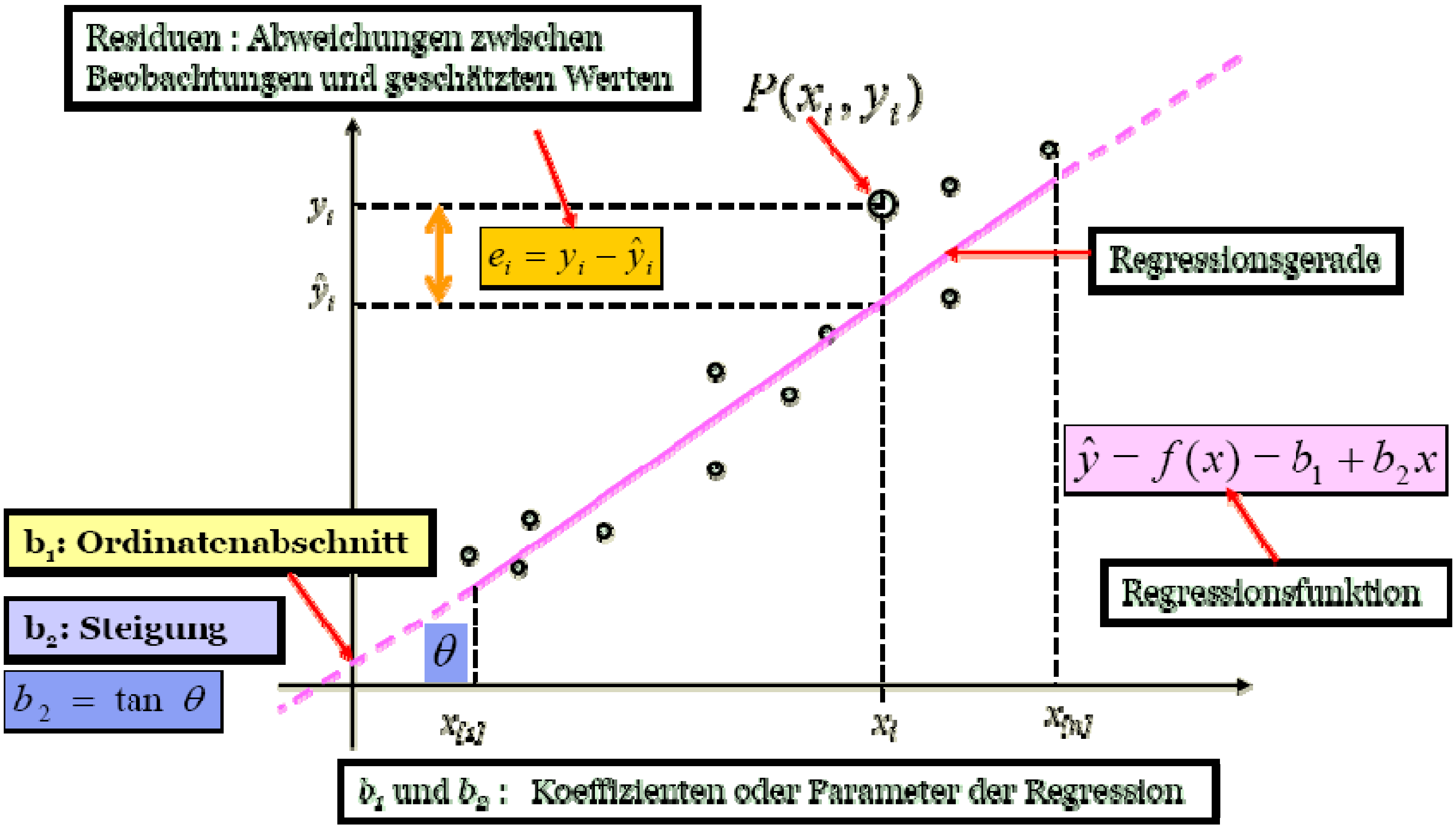
Nicht linearisierbar

Zusammenhang - modellieren

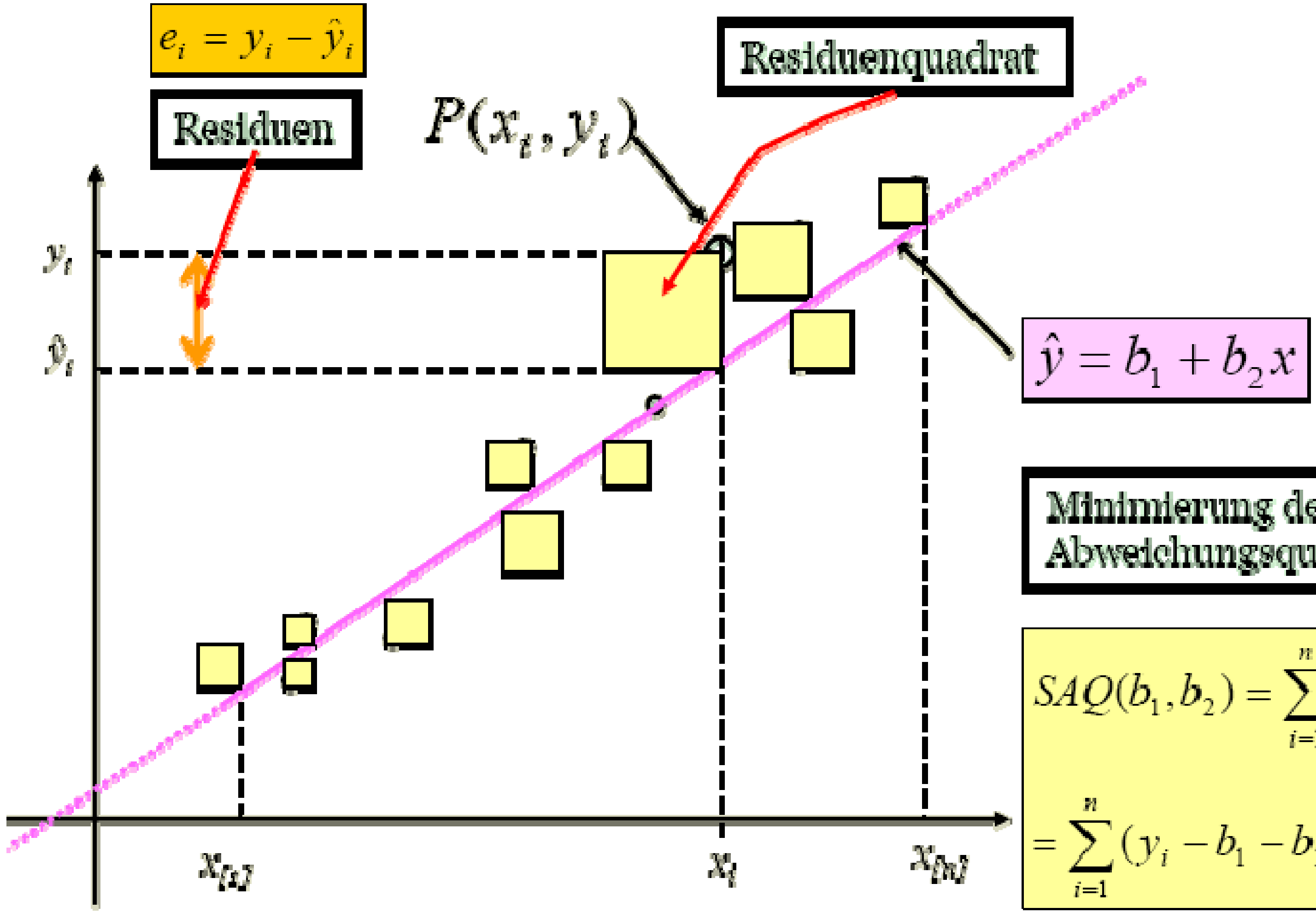


Unser Beispiel… Form? <u>Linear!</u>

Zusammenhang - modellieren



Zusammenhang - modellieren



Minimierung der Summe der Abweichungsquadrate *SAQ*

$$SAQ(b_1, b_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

Zusammenhang - modellieren

Zusammenhang **linear**

$Y = b_1 + b_2 X$

Achsen-abschnitt

Anstieg

Gr = *b* Gw + *a*

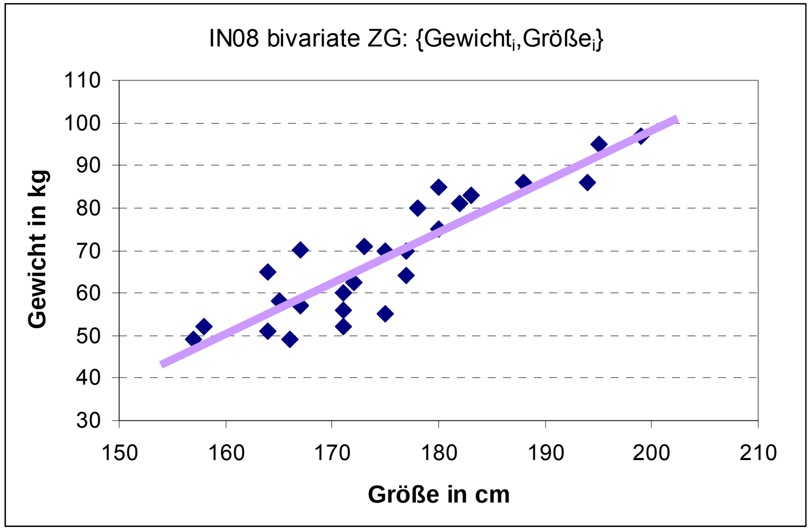
$$b_1 = \overline{Y} - b_2 \overline{X}$$

$$b_2 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Formel ergibt sich durch Minimieren der Abweichungsquadrate SAQ (Extremwertberechnung!!!)

Zusammenhang - Beispiel

Besteht ein Zusammenhang zwischen Wuchshöhe und Gewicht?



Berechne

Summe xi² = 826700

Summe yi² = 130342

Summe xi*yi = 324387

Weight	Height
64	177
70	177
80	178
75	180
85	180
81	182
83	183
86	188
86	194
95	195
97	199
49	157
52	158
51	164
65	164
58	165
49	166
57	167
57	167
70,1	167
52	171
56	171
60	171
62,5	172
71	173
55	175
70	175

Es fehlt noch
Summe xi = 4716
Summe yi = 1836,6

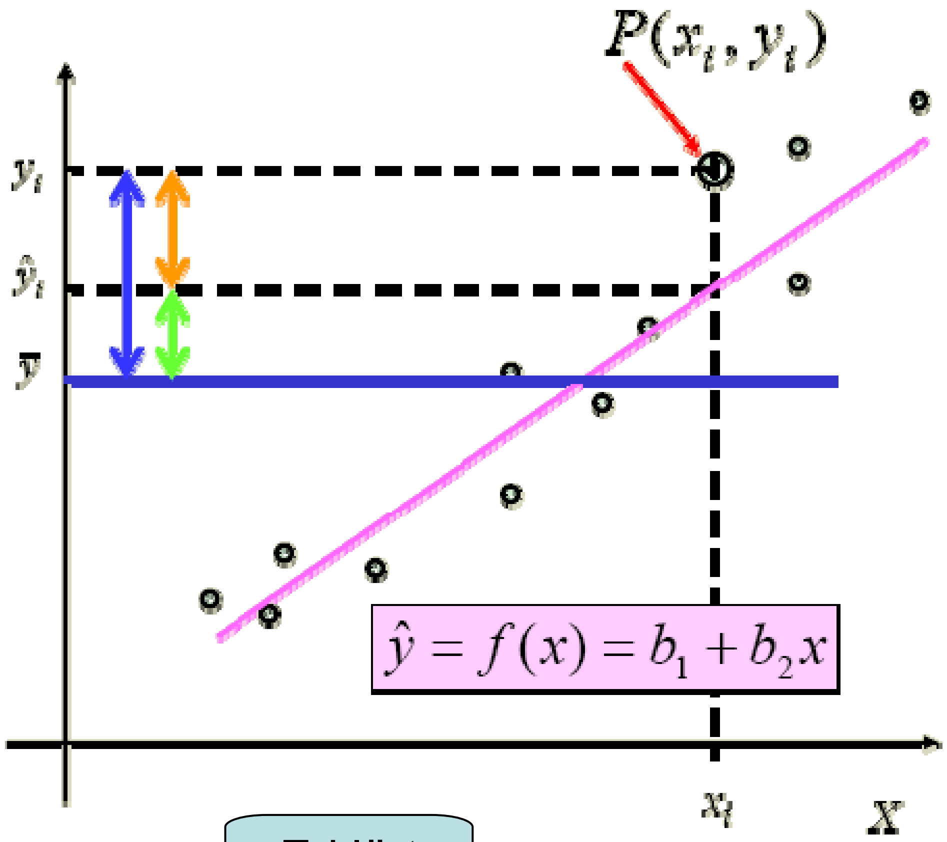
Es fehlt noch
Summe xi = 4716
Summe yi = 1836,6

Formel oben = 97035,3
Unten = 80244

b2 = 1,21

b1 = -142,2

Zusammenhang - modellieren



$y_i - \bar{y}$ — Zu erklärende Abweichung

$\hat{y}_i - \bar{y}$ — Durch das Modell erklärte Abweichung

$e_i = y_i - \hat{y}_i$ — Residuen oder nichterklärte Abweichungen

Bestimmtheitsmaß :

$$B = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{X;Y}^{\,2} * \text{Vorzeichen}(b_2)$$

SQE: Erklärt

SQT: Total

Zusammenhang - modellieren

Koeffizienten der Regressionsgerade sind **geschätzte** statistische Kenngrößen!!!
Daher… (siehe Formelsammlung S.24-25)

Konfidenzintervall:

b1

$$\text{für } \alpha: \quad \left[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\alpha}{2}}(n-2),\ \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\alpha}{2}}(n-2)\right]$$

$$\text{für } \beta: \quad \left[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\alpha}{2}}(n-2),\ \hat{\beta} + \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\alpha}{2}}(n-2)\right]$$

b2

Test mit T folgender verteilter PG:

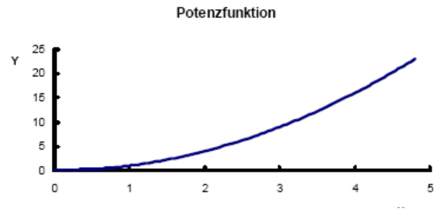
$$T_{\alpha_0} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} \quad \text{und} \quad T_{\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

z.B.
H0: $b_2 \geq b$
HA: $b_2 < b$

$$T_{\beta_0} < -t_{1-\alpha}(n-2)$$

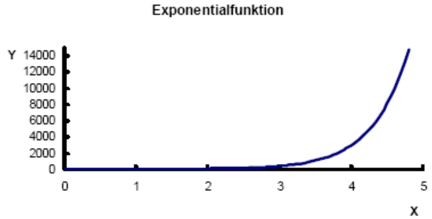
Linearisierung

$$\hat{y} = b_1 \cdot x^{b_2}$$



$$\boxed{\ln \hat{y} = \ln b_1 + b_2 \ln x} \Longleftrightarrow \boxed{\hat{y}^* = b^*_1 + b_2 x^*}$$

$$\hat{y} = b_1 \cdot e^{b_2 x}$$



$$\boxed{\ln \hat{y} = \ln b_1 + b_2 x} \Longleftrightarrow \boxed{\hat{y}^* - b^*_1 + b_2 x}$$

ABER:

Nicht zu jedem nichtlinearen Zusammenhang gibt es eine Linearisierung

thulke.statistics – leipzig – 0172-3449934 – ab @ thulke-statistics.de

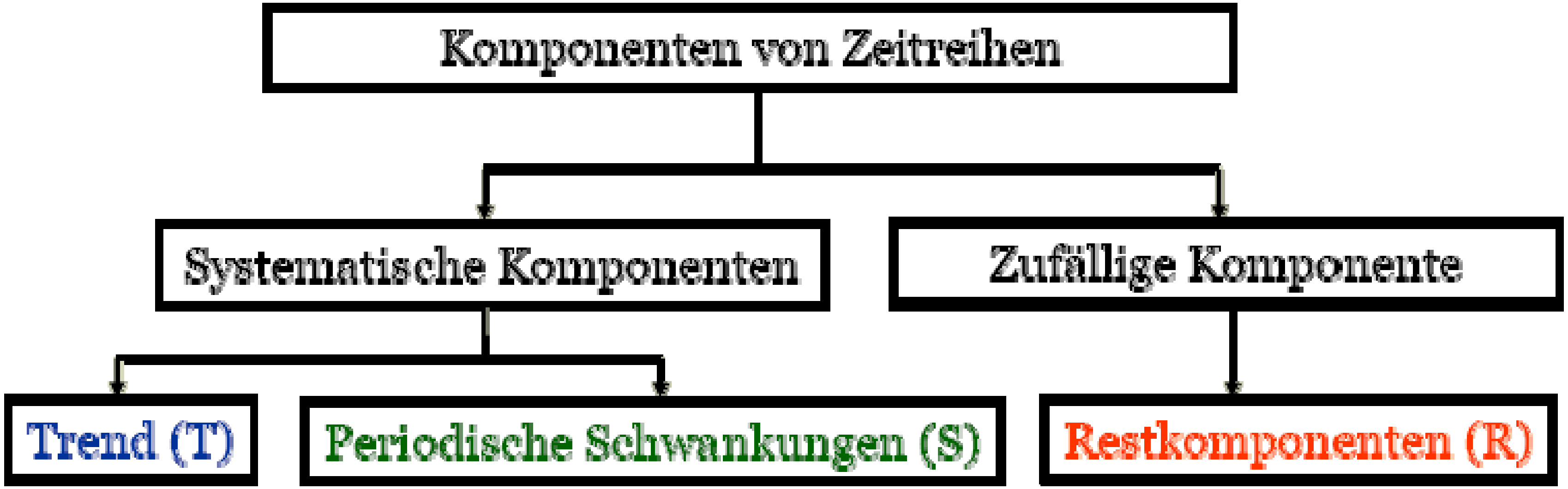
Modellieren von Zusammenhängen

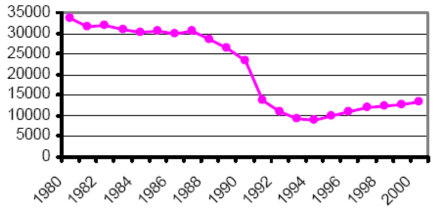
Multivariate Zufallsgrößen
Zusammenhangsmaße
Zusammenhangsmodelle
Trendbestimmung

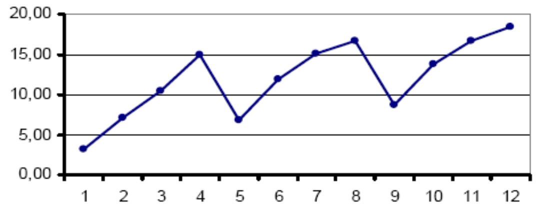
thulke.statistics – leipzig – 0172-3449934 – ab @ thulke-statistics.de

Zeitreihen

Bei einer Zeitreihe handelt es sich um eine Reihe von Werten ($y_1, y_2, \ldots, y_n$) eines Merkmals (Y), die zu verschiedenen Zeitpunkten oder verschiedenen Zeiträumen ($t=1, 2, \ldots, n$) erhoben werden.







thulke.statistics – leipzig – 0172-3449934 – ab @ thulke-statistics.de

Zeitreihen

Bestimmung der Trendkomponente einer Zeitreihe

1. Gleitender Durchschnitt der Länge g:

Für ungerade g mit g=2k+1:

$$\tilde{y}_t = \frac{1}{g} \sum_{j=-k}^{k} y_{t+j}$$

Für gerade g mit g=2k:

$$\tilde{y}_t = \frac{1}{g} \left(\frac{y_{t-k}}{2} + \sum_{j=-k+1}^{k-1} y_{t+j} + \frac{y_{t+k}}{2} \right)$$

Sehr einfach zu rechen aber verkürzt die Zeitreihe

Zeitreihen

Bestimmung der Trendkomponente einer Zeitreihe

2. Schätzen mit Regression:

„t statt x“

$$T_t = b_1 + b_2 \cdot t$$

$$b_1 = \bar{y} - b_2 \cdot \bar{t} \qquad b_2 = \frac{n \sum_{t=1}^{n} (t \cdot y_t) - \sum_{t=1}^{n} t \cdot \sum_{t=1}^{n} y_t}{n \sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t \right)^2}$$

Olle Kamellen…

thulke.statistics – leipzig – 0172-3449934 – ab @ thulke-statistics.de

Zeitreihen

Saisonwerte

$S_{kj}=Y_{kj}-T_{kj}$	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Gesamt	S_k
Quartal 1	-3,46	-3,68	-5,22	-4,66	-17,02	-4,25
Quartal 2	-0,74	0,56	-0,98	-0,82	-1,99	-0,50
Quartal 3	1,74	2,85	1,06	1,42	7,07	1,77
Quartal 4	5,33	3,59	1,85	1,16	11,93	2,98

$$S_k = \frac{1}{m}\sum_{j=1}^{m} s_{kj}$$

für k=1, ... , p mit p=4 und m=4

S_k: Durchschnittliche Saisonwert für das Quartal k

p: Anzahl der Quartale
m: Anzahl der Jahre

Modellieren von Zusammenhängen

Multivariate Zufallsgrößen

Zusammenhangsmaße

Zusammenhangsmodelle

Trendbestimmung

Statistik

- Daten erheben, verstehen, werten
- Hypothesen prüfen
- Modellieren von Zusammenhängen