Zusammenhänge zwischen metrischen Merkmalen

Darstellung des Zusammenhangs, Korrelation und Regression

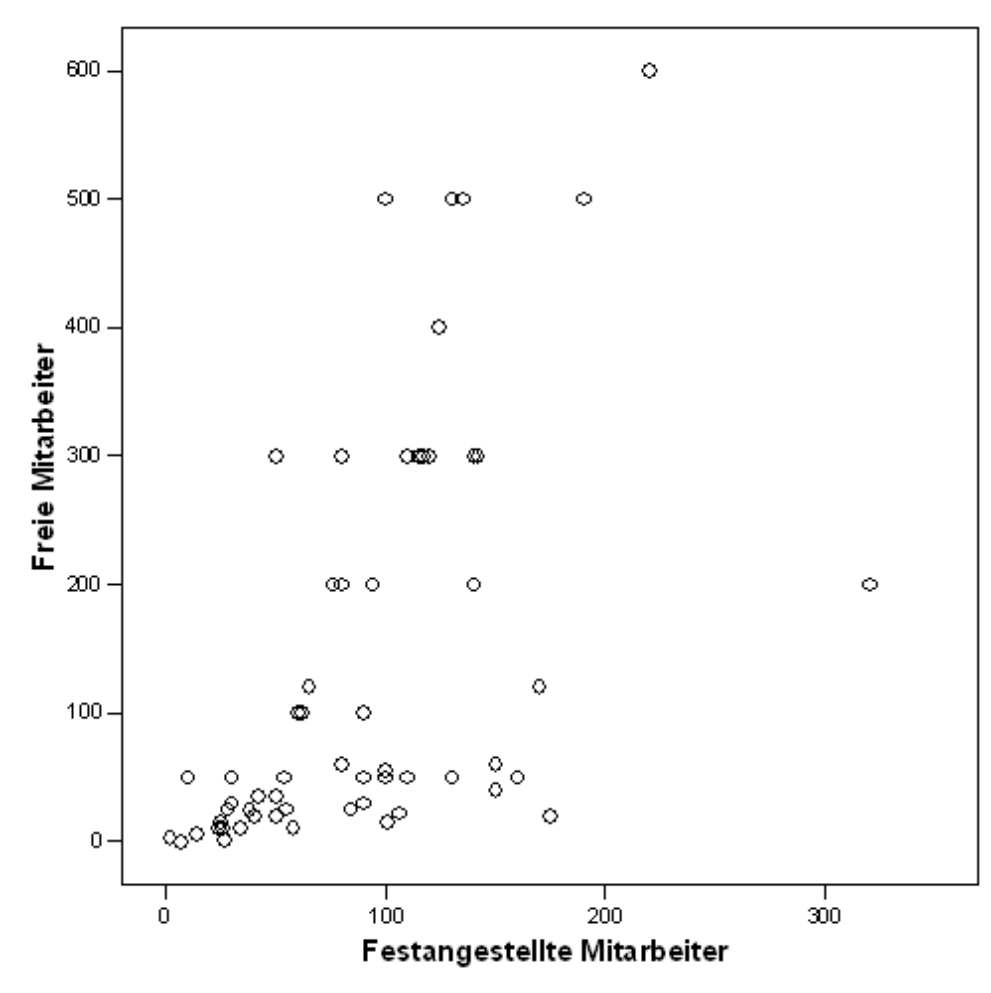
Daten liegen zu zwei metrischen Merkmalen vor:
Datenpaare $(x_i, y_i), \quad i = 1, \ldots, n$

Beispiel:
x: Anzahl der fest angestellten Mitarbeiter
y: Anzahl der freien Mitarbeiter

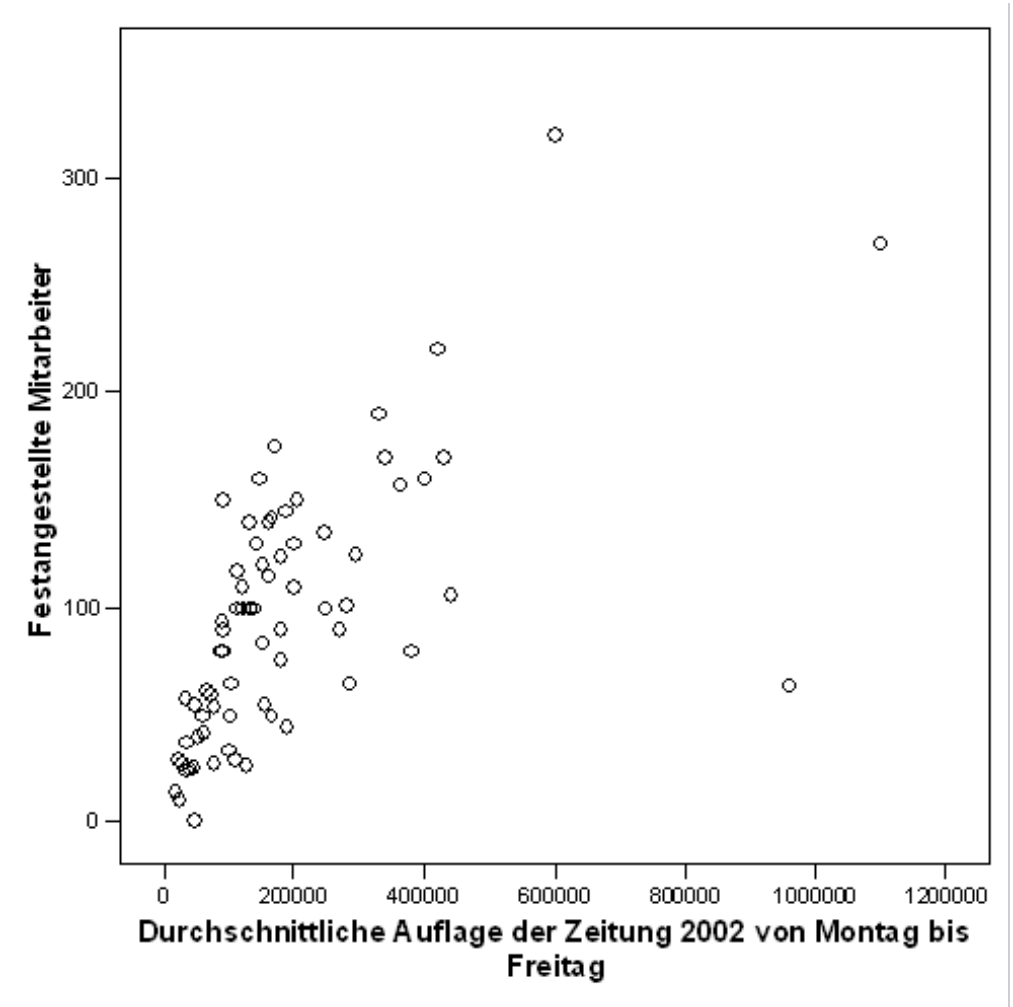
Frage:
Gibt es einen Zusammenhang zwischen diesen Merkmalen?
Wie lässt sich dieser Zusammenhang beschreiben?

Einfachste graphische Darstellung: Streudiagramm.
Die Datenpaare entsprechen Punkten in der Ebene („Punktwolke“)

Beispiel 1: Streudiagramm (mit SPSS)



Beispiel 2

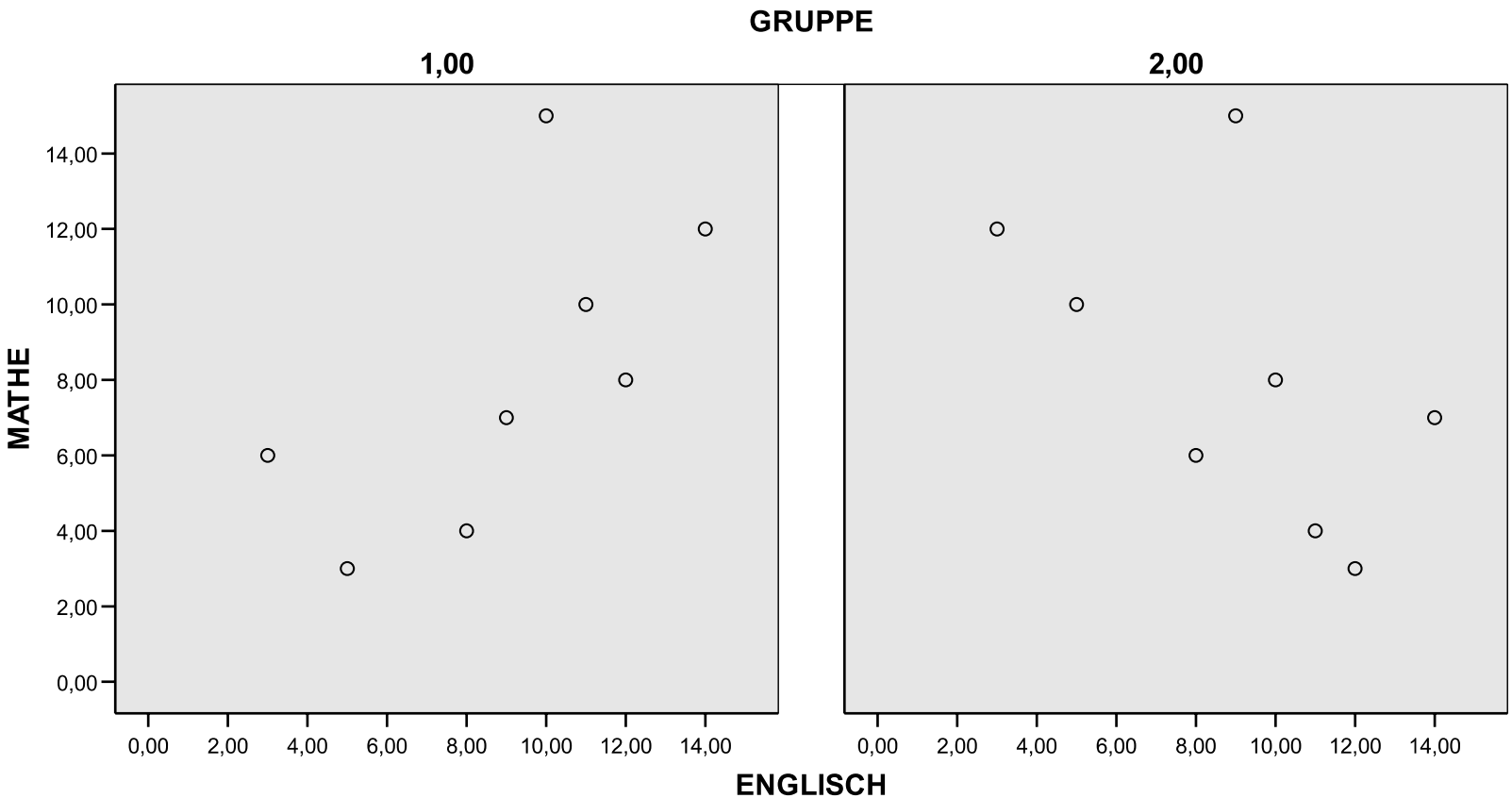

Festangestellte Mitarbeiter
300
200
100
0
0
200000
400000
600000
800000
1000000
1200000
Durchschnittliche Auflage der Zeitung 2002 von Montag bis Freitag

Beispiel 3

Punkte in Englisch und Mathematik

Schüler	Gruppe 1		Gruppe 2	
	Englisch	Mathe	Englisch	Mathe
1	14	12	10	8
2	9	7	8	6
3	5	3	3	12
4	3	6	5	10
5	11	10	14	7
6	8	4	9	15
7	10	15	11	4
8	12	8	12	3
Mittelwert	9.0	8.1	9.0	8.1
Standardabweichung	3.6	4.1	3.6	4.1

Beispiel 3 (Streudiagramme)



Kovarianz

Maß für den Zusammenhang der beiden Merkmale:

Daten: $(x_i, y_i), \quad i = 1, \ldots, n$

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

Beachte:

- Summand i positiv, falls x_i und y_i relativ zum Mittelwert das gleiche Vorzeichen haben.
- Für s_{xx} ergibt sich die Varianz von X.
- Die Kovarianz hängt sowohl von der Streuung als auch von dem Zusammenhang der beiden Merkmale ab.

Bravais-Pearson-Korrelationskoeffizient

Der Bravais-Pearson-Korrelationskoeffizient ergibt sich aus den Daten $(x_i, y_i), i = 1, \ldots, n$ durch

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{S_{xy}}{S_x S_y}$$

Wertebereich: $-1 \leq r \leq 1$

$r > 0$	positive Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
$r < 0$	negative Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
$r = 0$	keine Korrelation, unkorreliert, kein linearer Zusammenhang

Punkte in Englisch und Mathematik

Gruppe 1:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{9.57}{3.641} = 0.65$$

Gruppe 2:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-8.29}{3.6 \cdot 4.1} = -0.56$$

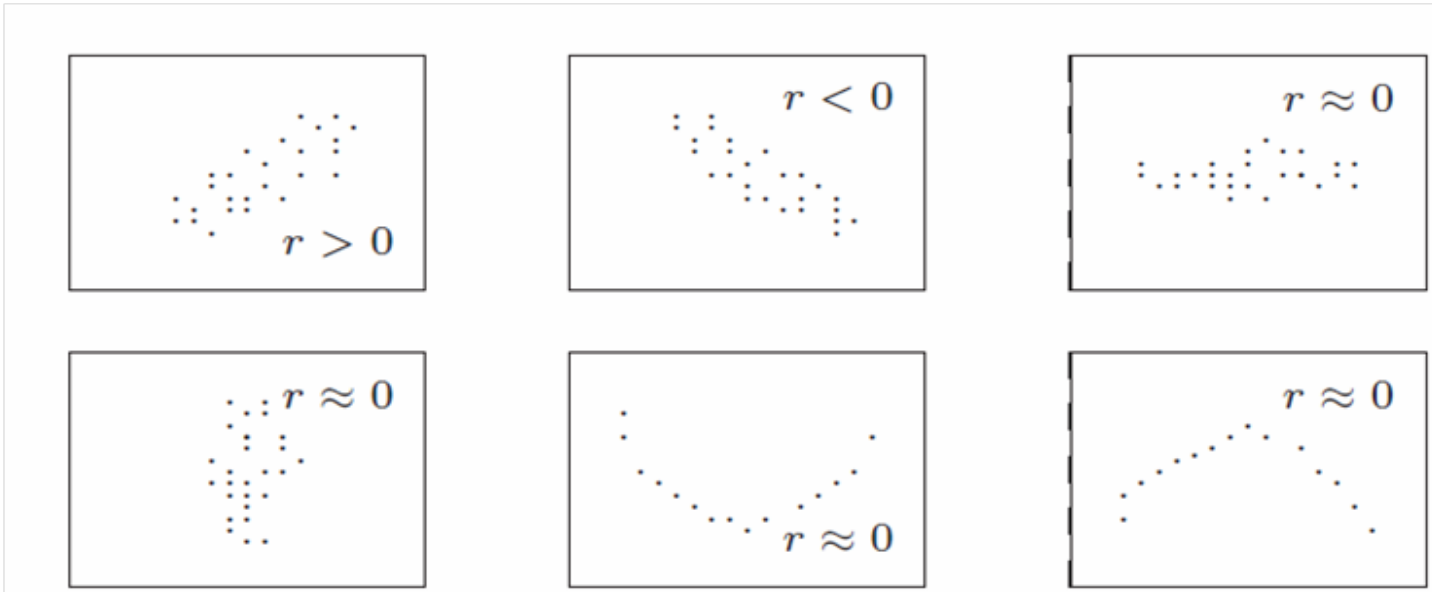
Gruppe 1: positiver linearer Zusammenhang
Gruppe 2: negativer linearer Zusammenhang

Eigenschaften des Korrelationskoeffizienten

- Maß für den linearen Zusammenhang
- Ändert sich nicht bei linearen Transformationen
- Symmetrisch (Korrelation zwischen x und y = Korrelation zwischen y und x)
- Positive Korrelation bedeutet: Je größer x, desto größer im Durchschnitt y
- Korrelation = +1 oder -1, falls die Punkte genau auf einer Geraden liegen
- Korrelation = 0 bedeutet keinen linearen Zusammenhang, aber nicht Unabhängigkeit
- Korrelation empfindlich gegenüber Ausreißern

Eigenschaften von r

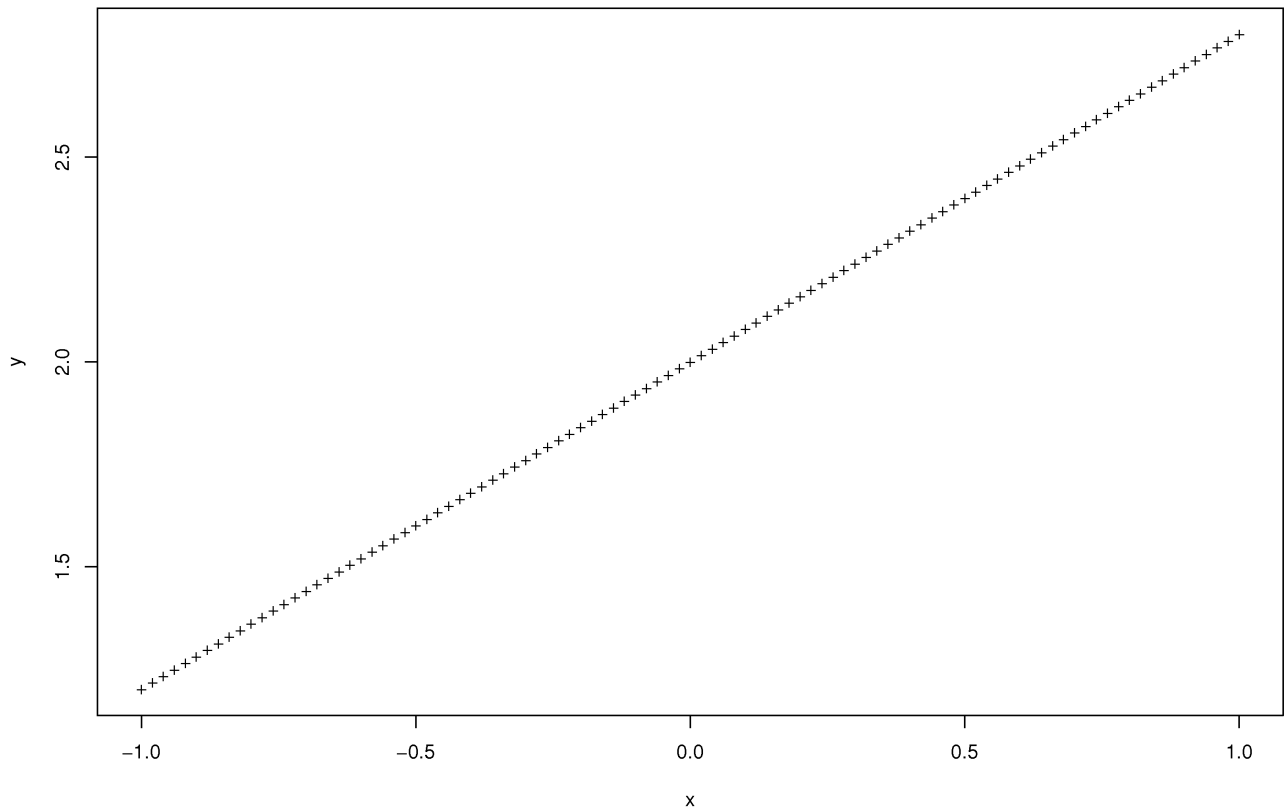
- r misst Stärke des *linearen* Zusammenhangs.



Punktkonfigurationen und Korrelationskoeffizienten (qualitativ)

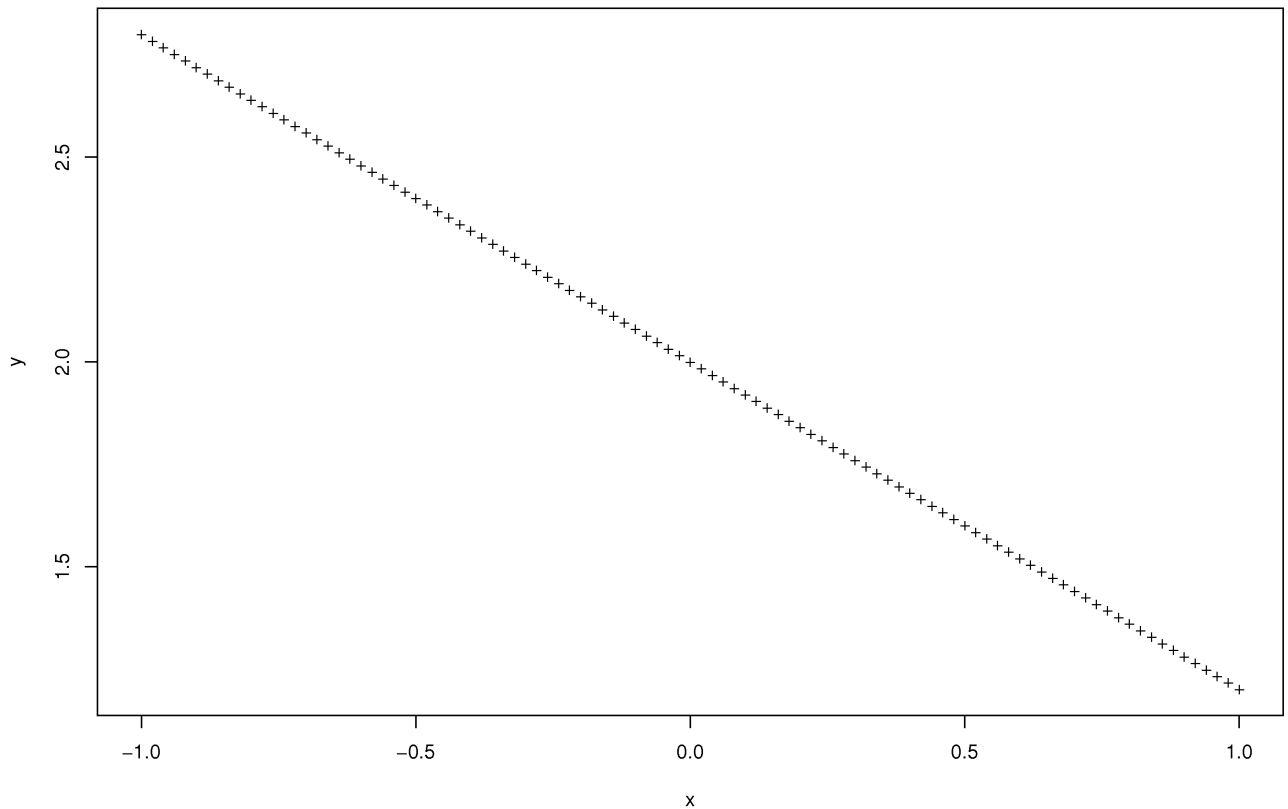
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 1: Lineare (unverrauschte) Funktion, $y = 0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



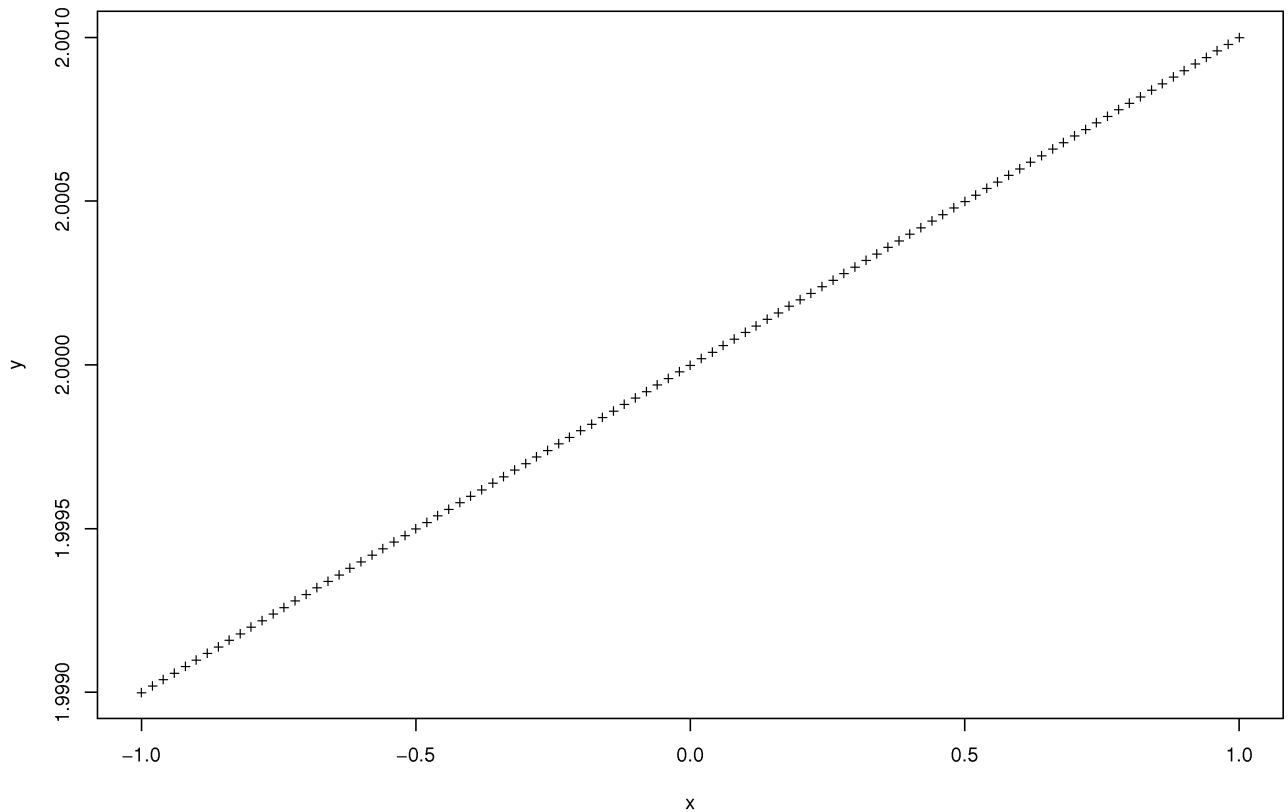
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 2: Lineare (unverrauschte) Funktion, $y = -0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



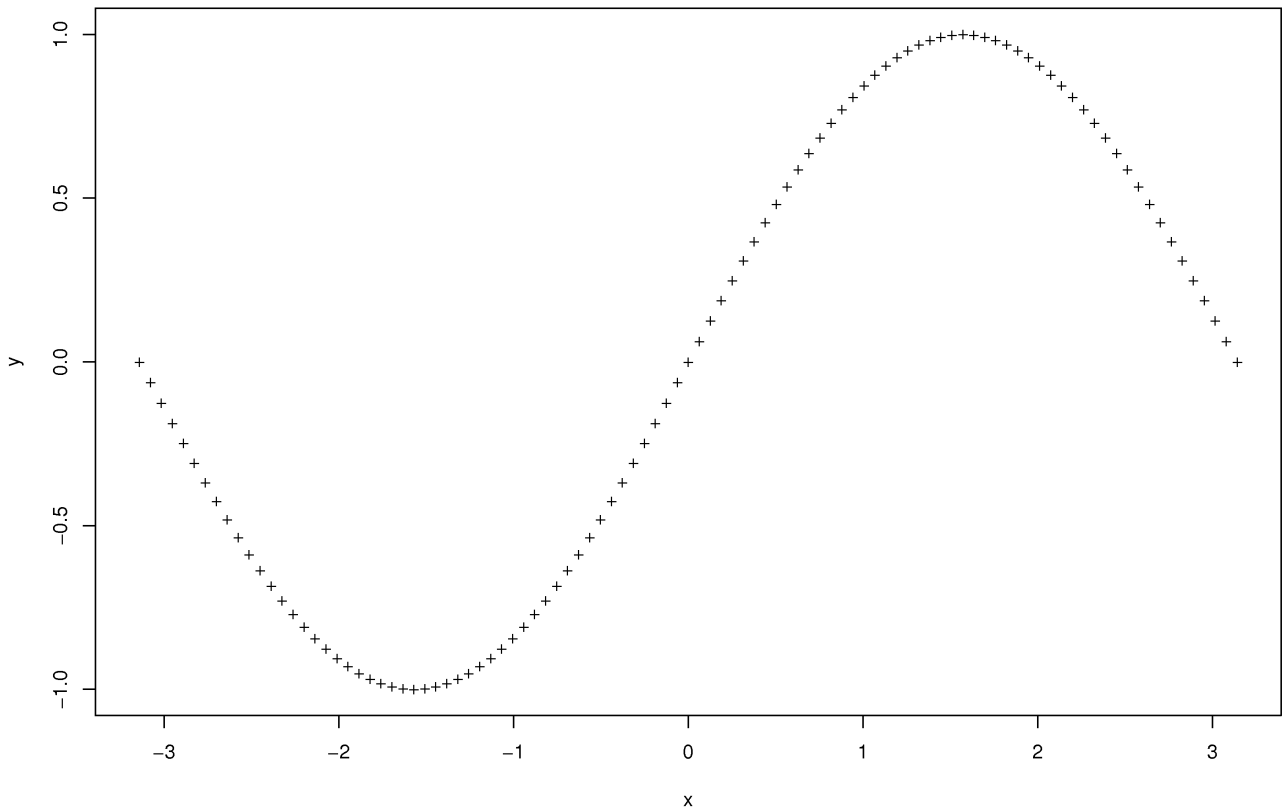
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 3: Lineare (unverrauschte) Funktion, $y = 0.001x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



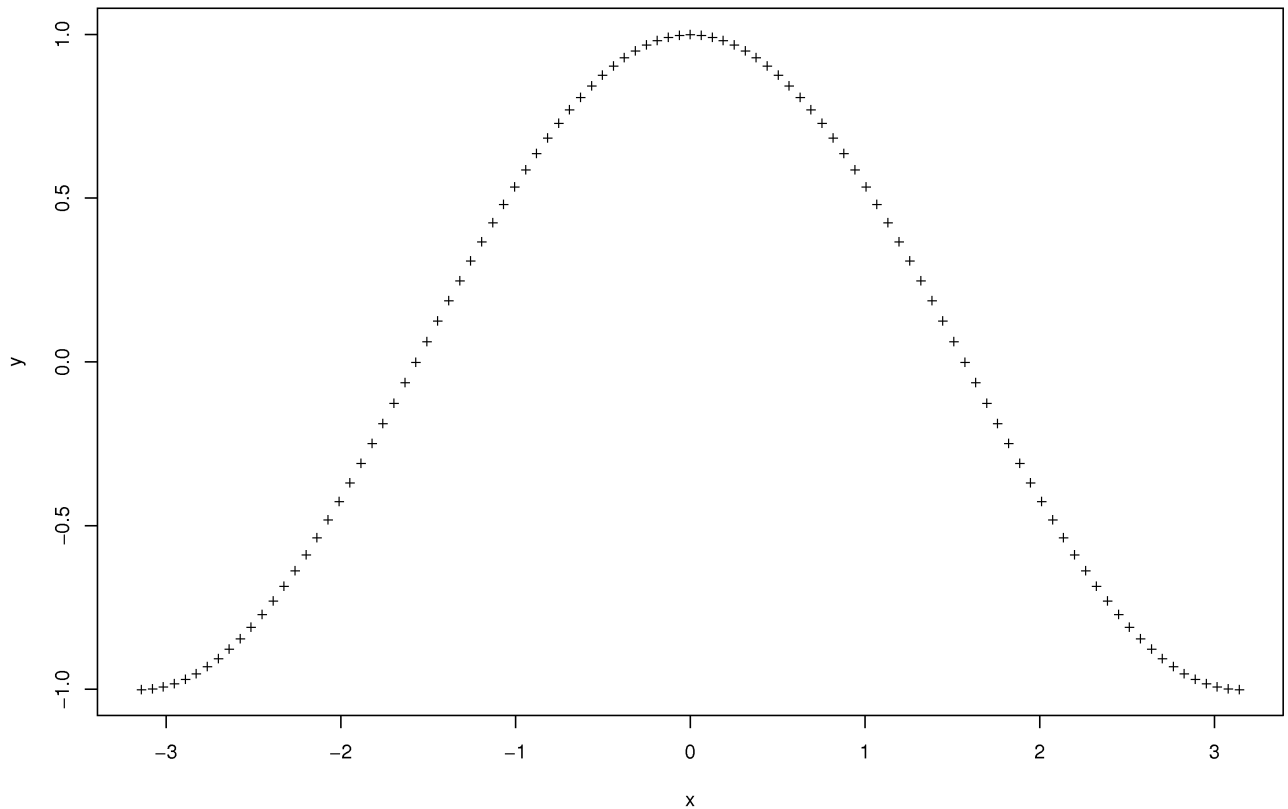
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 4: Periodische (unverrauschte) Funktion, $y = \sin(x)$, 101 equidistante Stützstellen im Intervall $[-\pi, \pi]$, $r =$



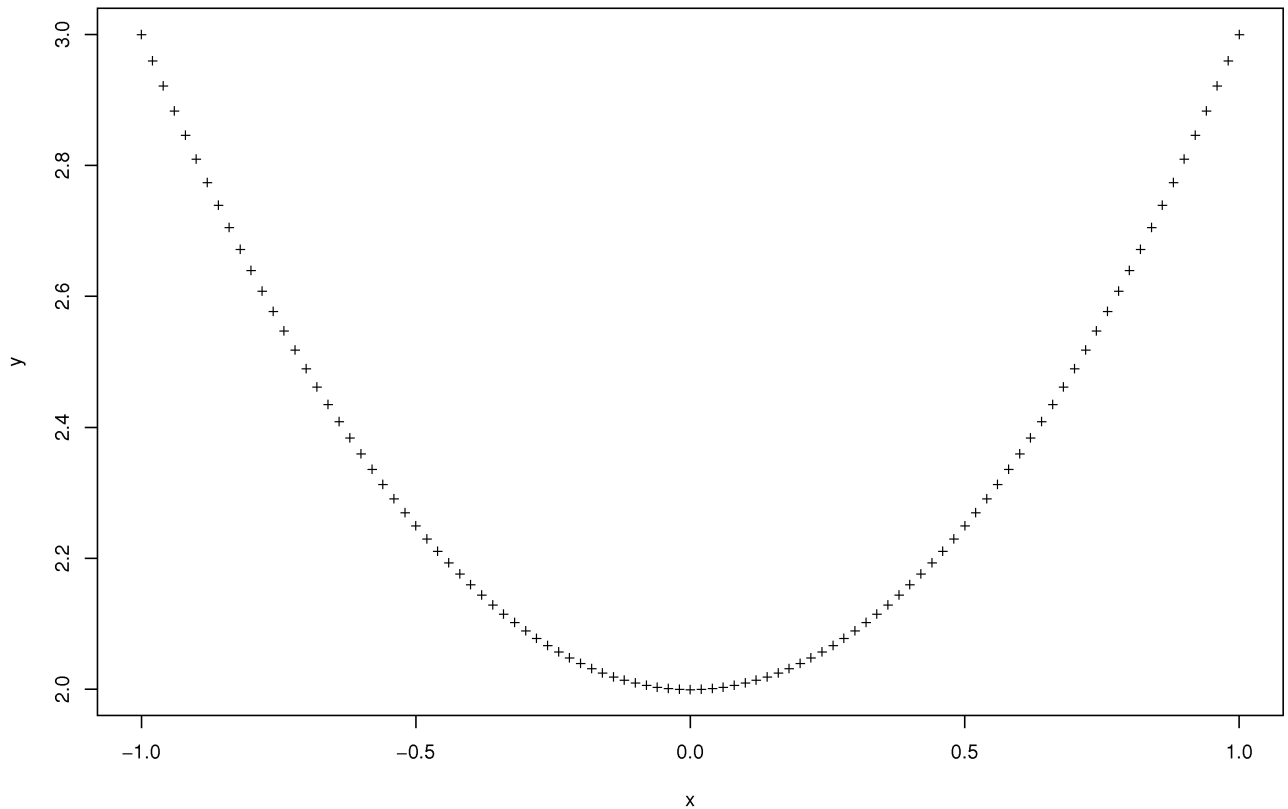
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 5: Periodische (unverrauschte) Funktion, $y = \cos(x)$, 101 equidistante Stützstellen im Intervall $[-\pi, \pi]$, $r =$



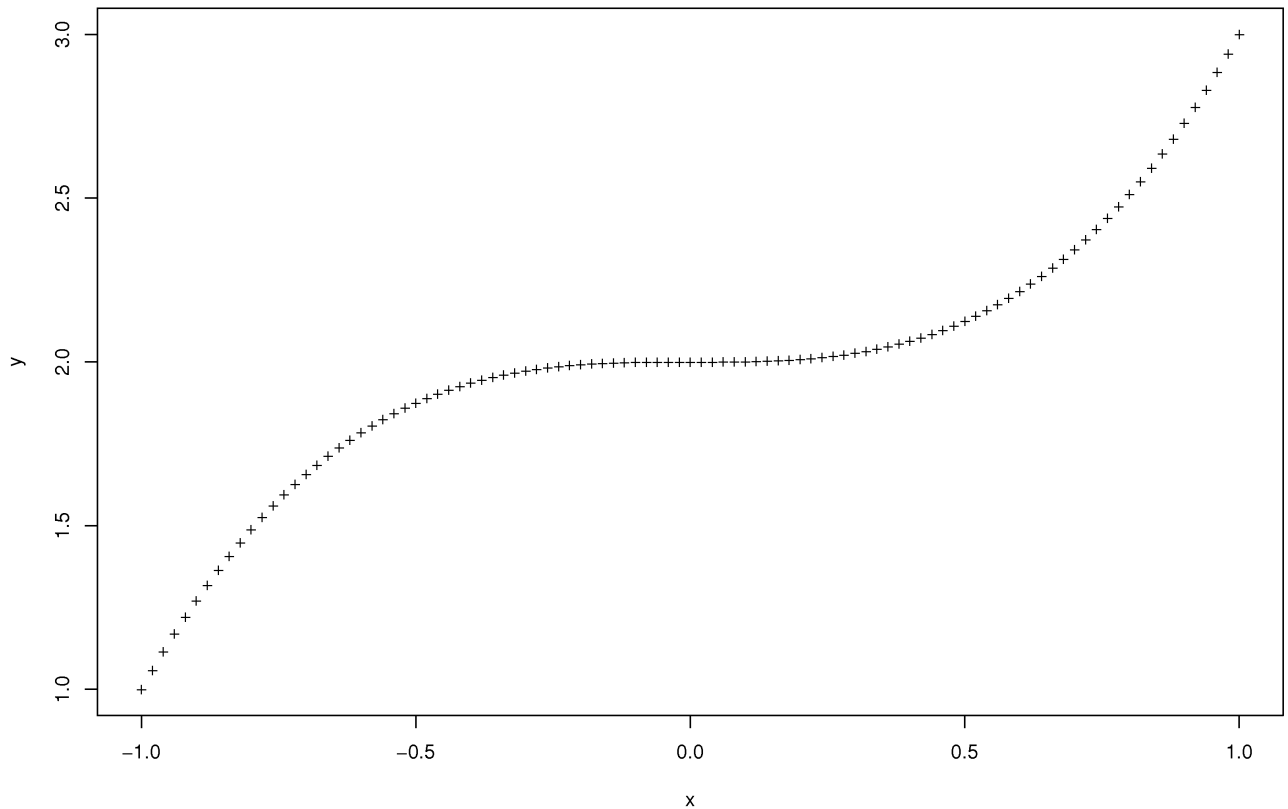
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 6: Quadratische (unverrauschte) Funktion, $y = x^2 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r =$



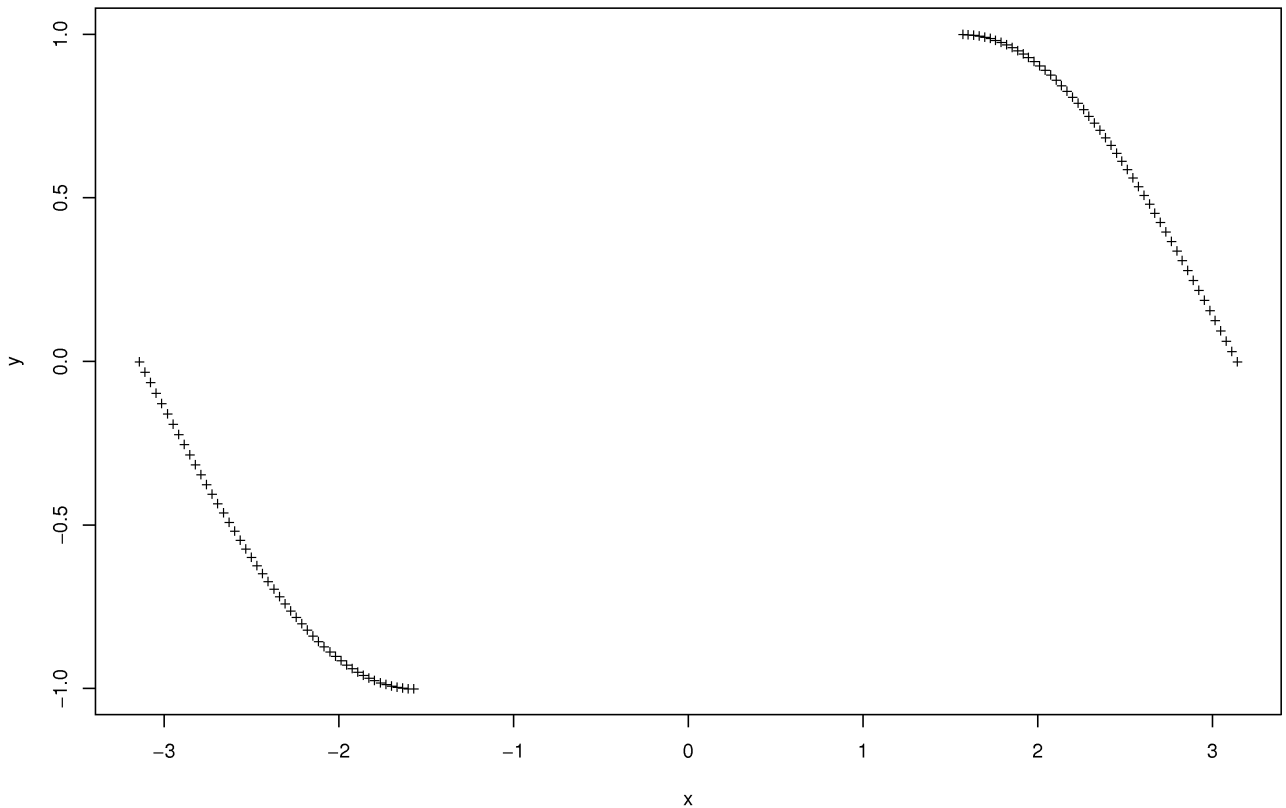
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 7: Kubische (unverrauschte) Funktion, $y = x^3 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r =$



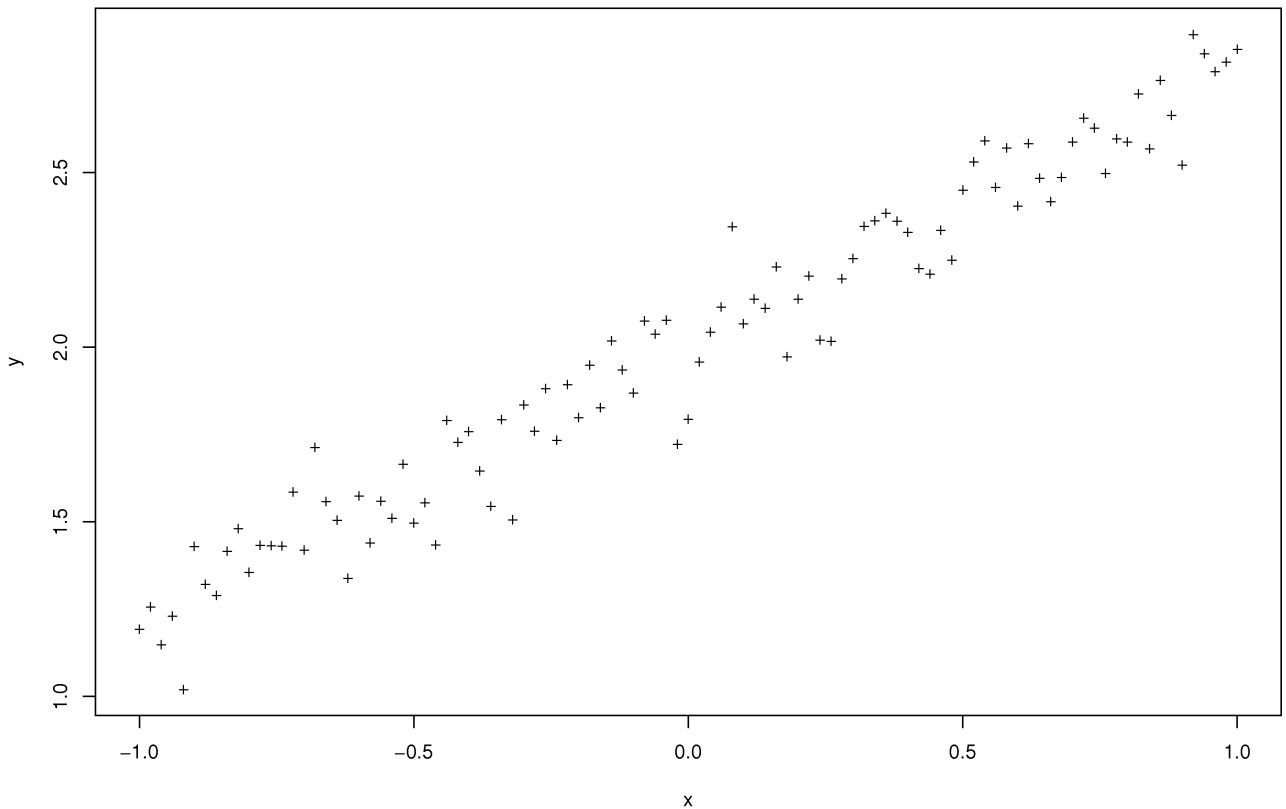
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 8: Abschnittweise definierte (unverrauschte) Funktion $y = \sin(x)$, 50 und 51 equidistante Stützstellen in den Intervallen $[-\pi, -\frac{\pi}{2}]$ und $[\frac{\pi}{2}, \pi]$, $r =$



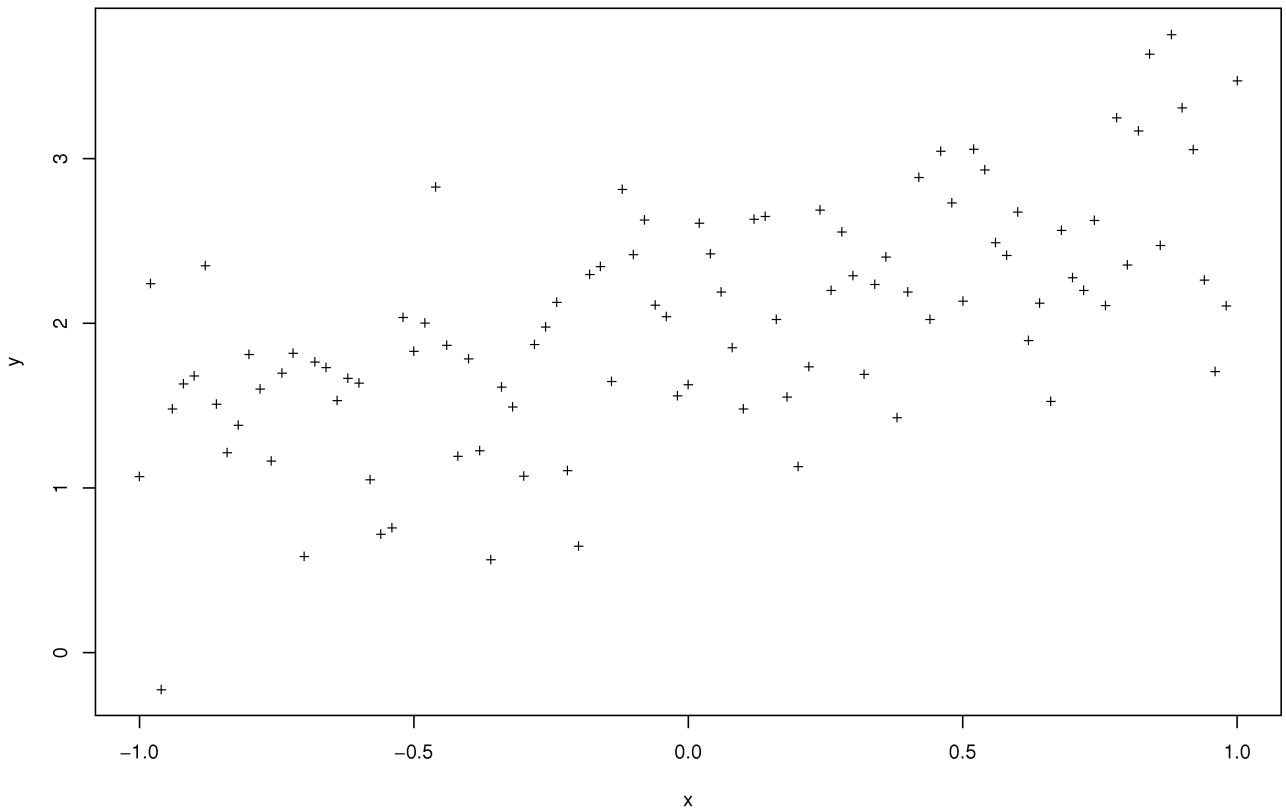
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 9: Lineare, schwach verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



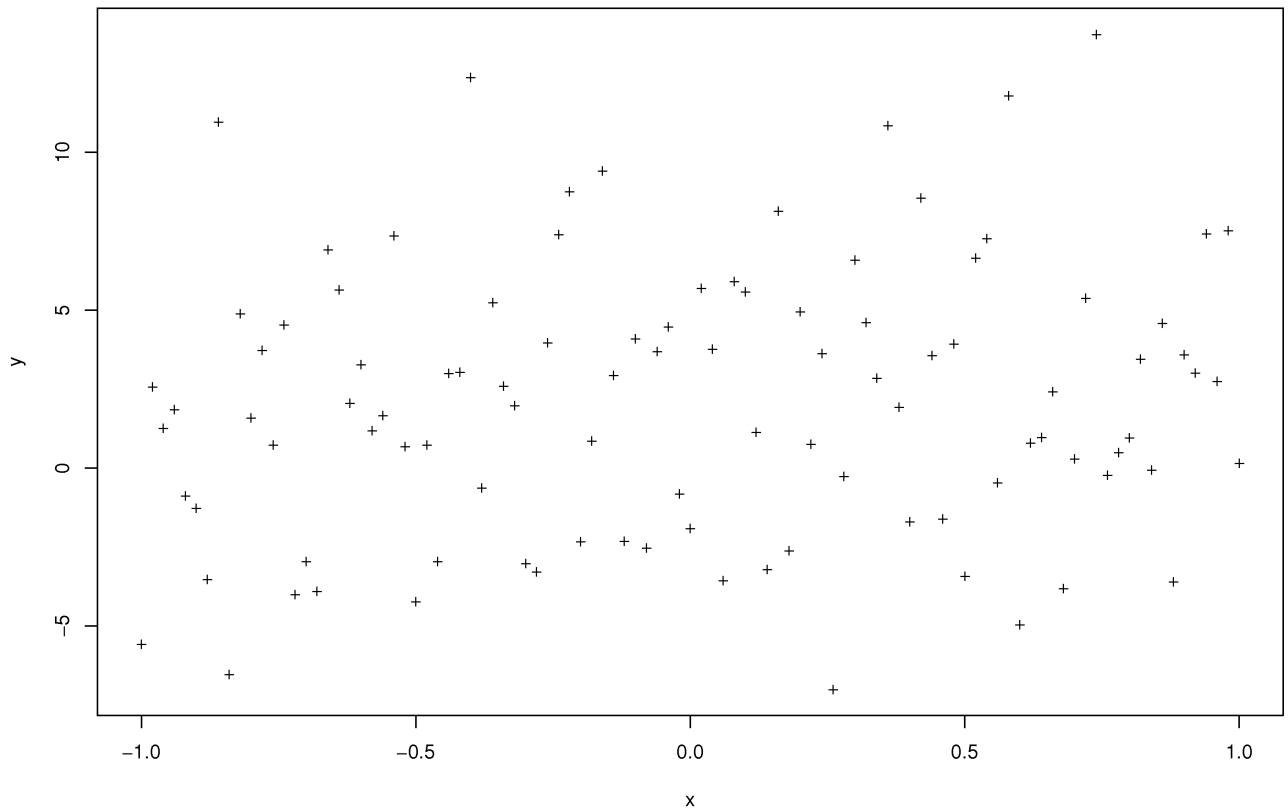
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 10: Lineare, stärker verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 0.5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



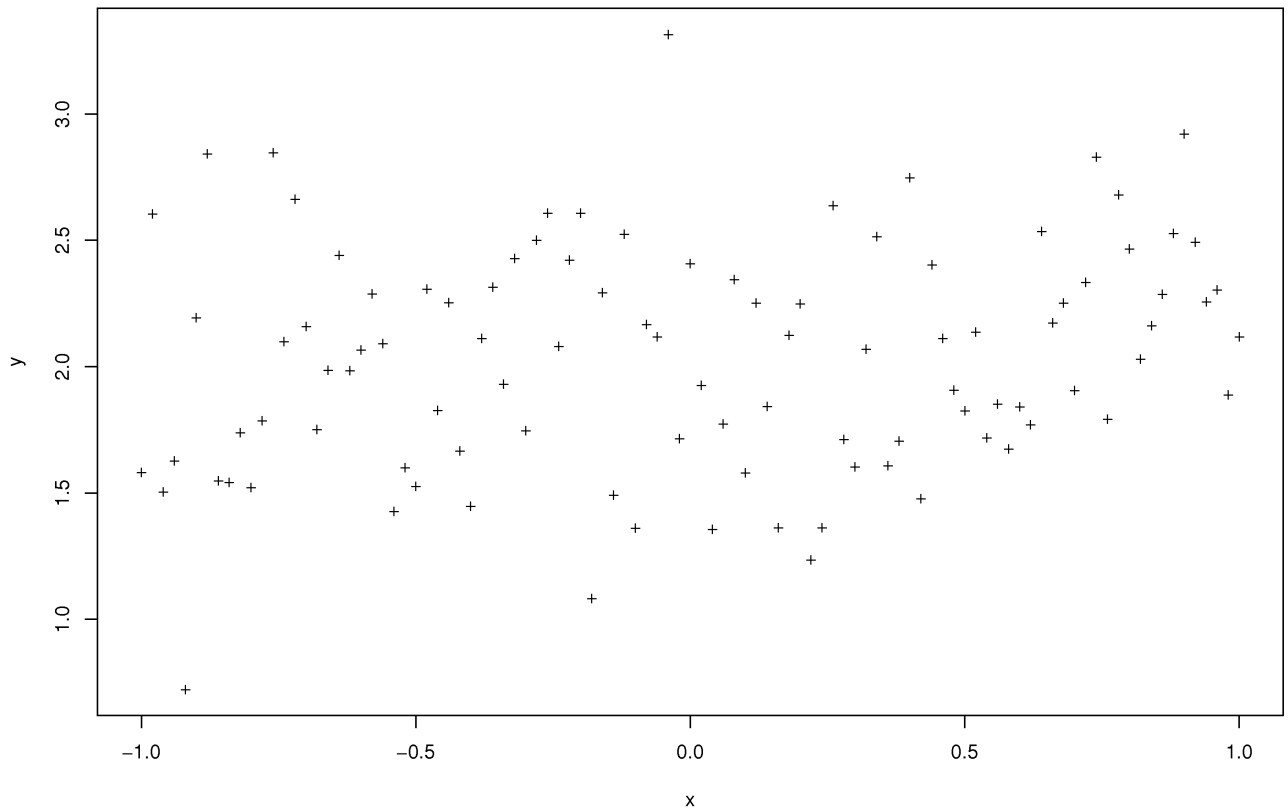
Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 11: Lineare, stark verrauschte Funktion, $y = 0.8x + 2.0 + N(0, 5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Einige Beispiele von exakten und verrauschten Zusammenhängen

Beispiel 12: Lineare, stärker verrauschte Funktion, $y = 0.1x + 2.0 + N(0, 0.5)$, 101 equidistante Stützstellen im Intervall [-1,1], $r =$



Lineare Transformationen

- Bei exakten lineare Zusammenhängen gilt:

$$r = +1 \text{ bzw. } -1 \quad \Leftrightarrow \quad Y = aX + b \text{ mit } b > 0 \text{ bzw. } b < 0$$

- Lineare Transformationen

$\tilde{X} = a_X X + b_X, \tilde{Y} = a_Y Y + b_Y, a_X, a_Y \neq 0$
r Korrelationskoeffizient zwischen X und Y
$\tilde{r}$ Korrelationskoeffizient zwischen $\tilde{X}$ und $\tilde{Y}$

$$\Rightarrow \begin{array}{lll} \tilde{r} = r & \Leftrightarrow & a_X, a_Y > 0 \text{ oder } a_X, a_Y < 0 \\ \tilde{r} = -r & \Leftrightarrow & a_X > 0, a_Y < 0 \text{ oder } a_X < 0, a_Y > 0. \end{array}$$

Vektor-Darstellung

Definiere die zentrierten Datenvektoren

$$x_Z = (x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_n - \bar{x})'$$
$$y_Z = (y_1 - \bar{y}, \ldots, y_i - \bar{y}, \ldots, y_n - \bar{y})'$$

$\Rightarrow r = \frac{x_Z' y_Z}{||x_Z||\,||\bar{y}_Z||}$, mit $||\,.\,||$ euklidische Norm.

Aus der Schwarz-Cauchy-Ungleichung folgt

$$|x_Z' y_Z| \leq ||x_Z||\,||y_Z||,$$

d.h. $-1 \leq r \leq +1$.

Spearmans Korrelationskoeffizient = Rang-Korrelationskoeffizient

X, Y (mindestens) ordinal

Idee: Gehe von Werten $x_i, \quad i = 1, \ldots, n$ und $y_i, \quad i = 1, \ldots, n$ über zu ihren Rängen.

$$x_{(1)} \leq \ldots x_{(i)} \ldots \leq x_{(n)}$$

$$rg(x_{(i)}) = i,$$

analog für $y_{(1)}, \ldots, y_{(n)}$.

Beispiel

x_i	2.3	7.1	1.0	2.1
$rg(x_i)$	3	4	1	2

bei Bindungen (ties):

x_i	2.3	7.1	1.0	2.1	2.3
	3.5	5	1	2	3.5

$\Rightarrow$ Durchschnittsrang $\frac{3+4}{2} = 3.5$ vergeben.
Also: Urliste der Größe nach durchsortieren
$\Rightarrow$ Ranglisten $rg(x_i), rg(y_i), i = 1, \ldots, n$ vergeben (bei ties: Durchschnittsränge)
Idee: Berechne den Korrelationskoeffizienten nach Bravais-Pearson für die Ränge statt für die Urliste.

Definition: Spearmans Korrelationskoeffizient

Der *Korrelationskoeffizient nach Spearman* ist definiert durch

$$r_{SP} = \frac{\sum(rg(x_i) - \bar{rg}_X)(rg(y_i) - \bar{rg}_Y)}{\sqrt{\sum(rg(x_i) - \bar{rg}_X)^2 \sum(rg(y_i) - \bar{rg}_Y)^2}}.$$

Wertebereich: $-1 \leq r_{SP} \leq 1$

Interpretation

$r_{SP} > 0$ gleichsinniger monotoner Zusammenhang,

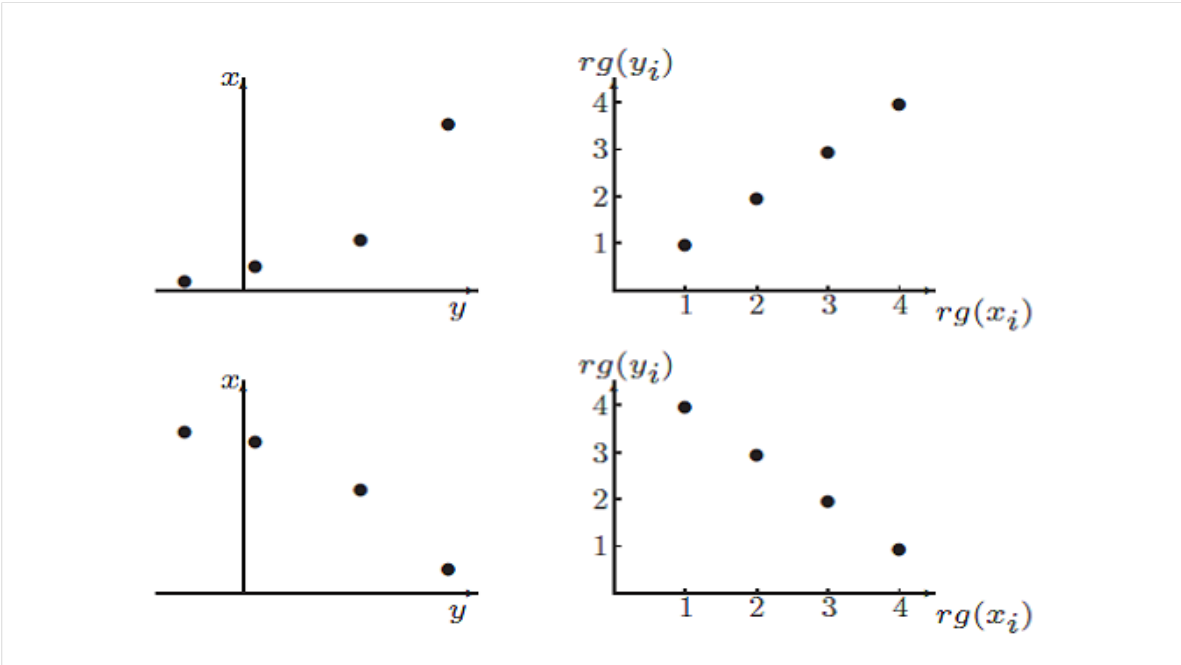
Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein

$r_{SP} < 0$ gegensinniger monotoner Zusammenhang,

Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß

$r_{SP} \approx 0$ kein monotoner Zusammenhang

Extremfälle



Extremfälle für Spearmans Korrelationskoeffizienten, $r_{SP} = 1$ (oben) und $r_{SP} = -1$ (unten)

Spearmans Korrelationskoeffizient misst monotone (auch nichtlineare) Zusammenhänge!

Bemerkungen:

- Rechentechnische Vereinfachungen:

$$\bar{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2,$$

$$\bar{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2.$$

Rechentechnisch günstige Version von r_{SP}:

Daten: $(x_i, y_i),\ i = 1, \ldots, n, \quad x_i \neq x_j,\ y_i \neq y_j$ für alle i, j
Rangdifferenzen: $d_i = rg(x_i) - rg(y_i)$

$$r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n}$$

Voraussetzung: keine Bindungen

Monotone Transformationen

$\tilde{X} = g(X)$ $\quad g$ streng monoton,
$\tilde{Y} = h(Y)$ $\quad h$ streng monoton

$\Rightarrow \quad r_{SP}(\tilde{X}, \tilde{Y}) = r_{SP}(X, Y)$,
wenn g *und* h monoton wachsend
bzw. g *und* h monoton fallend sind,

$r_{SP}(\tilde{X}, \tilde{Y}) = -r_{SP}(X, Y)$,
wenn g monoton wachsend *und* h
monoton fallend bzw. g monoton
fallend *und* h monoton wachsend sind.

Kendall´s Tau

Betrachte Paare von Beobachtungen (x_i, y_i) und (x_j, y_j)

Ein Paar heißt:

konkordant, falls $x_i < x_j$ und $y_i < y_j$
oder $x_i > x_j$ und $y_i > y_j$

diskordant, falls $x_i < x_j$ und $y_i > y_j$
oder $x_i > x_j$ und $y_i < y_j$

N_C: Anzahl der konkordanten Paare
N_D: Anzahl der diskordanten Paare

$$\tau_a = \frac{N_C - N_D}{n(n-1)/2}$$

Kendall´s Tau

Andere Varianten

- Goodman & Kruskal γ-Koeffizient

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

- Somers D wird typischerweise verwendet wenn Y binär ist
  T_x: Anzahl der Paare mit ungleichem y und gleichem x
  („Ties“ = Bindungen)

$$D_{xy} := \frac{N_C - N_D}{N_C + N_D + T_x} = \frac{N_C - N_D}{\text{Anzahl Paare mit ungleichem y}}$$

Kendall´s τ, Spearman´s r_{sp}

Beispiel:

					τ	r_{sp}
rg X	1	2	3	4	0.33	0.6
rg Y	2	1	4	3		
rg X	1	2	3	4	0.33	0.4
rg Y	1	3	4	2		

r_{sp} bestraft Abweichung stärker als τ

Unterschiede Kendall´s τ, Spearman´s ρ

- ρ verwendet Abstände auf der Rang-Skala
- τ orientiert sich an Paarvergleichen
- τ hat theoretische Entsprechung
- τ in der Regel kleiner als ρ

Dichotome und stetige Merkmale: Punktbiseriale Korrelation

Korrelations-Koeffizient zwischen dichotomen und metrischem Merkmal

$X \in \{0, 1\}$ Y metrisch

$$r_{XY} = \frac{\bar{Y}_1 - \bar{Y}_0}{\widetilde{S}_Y} \cdot \sqrt{\frac{n_0 n_1}{N^2}}$$

$\bar{Y}_0$ Mittelwert bei $X = 0$,
$\bar{Y}_1$ Mittelwert bei $X = 1$

Entspricht normiertem Abstand der Gruppenmittelwerte.

Dichotome und stetige Merkmale

- Beispiel 1 Kredit Scoring: Die Kreditwürdigkeit wird mit einem Scorewert gemessen (Schufa score)
  Dieser Scorewert soll auf seine Prognosegüte geprüft werden
  Variable : Y=1 (Eintrag nach 1.5 Jahren (Default) Y=0 kein Eintrag
- Beispiel 2: Blutserum Konzentration und stress-induzierte Herzinfarkte
  X: Marker für Herzinfarkt und
  Y: Infarkt während der WM (Gruppen)

ROC-Kurve

Jetzt Y dichotome Zielgröße und X metrische Einflussgröße:
$Y = 1 \longrightarrow$ Ausfall (krank)
$Y = 0 \longrightarrow$ kein Ausfall (gesund)
In der medizinischen Literatur ist das Testergebnis m:

$$\hat{Y}_i = 1 \Leftrightarrow x_i \geq c$$

Sensitivität und Spezifität

Richtig Positiv = Sensitivität:

$$f(\hat{Y} = 1|Y = 1) \quad = \quad f(x \geq c|Y = 1) = S_1(c)$$

$S_1(c)$ stellt die Survivorfunktion dar.

Richtig negativ = Spezifität:

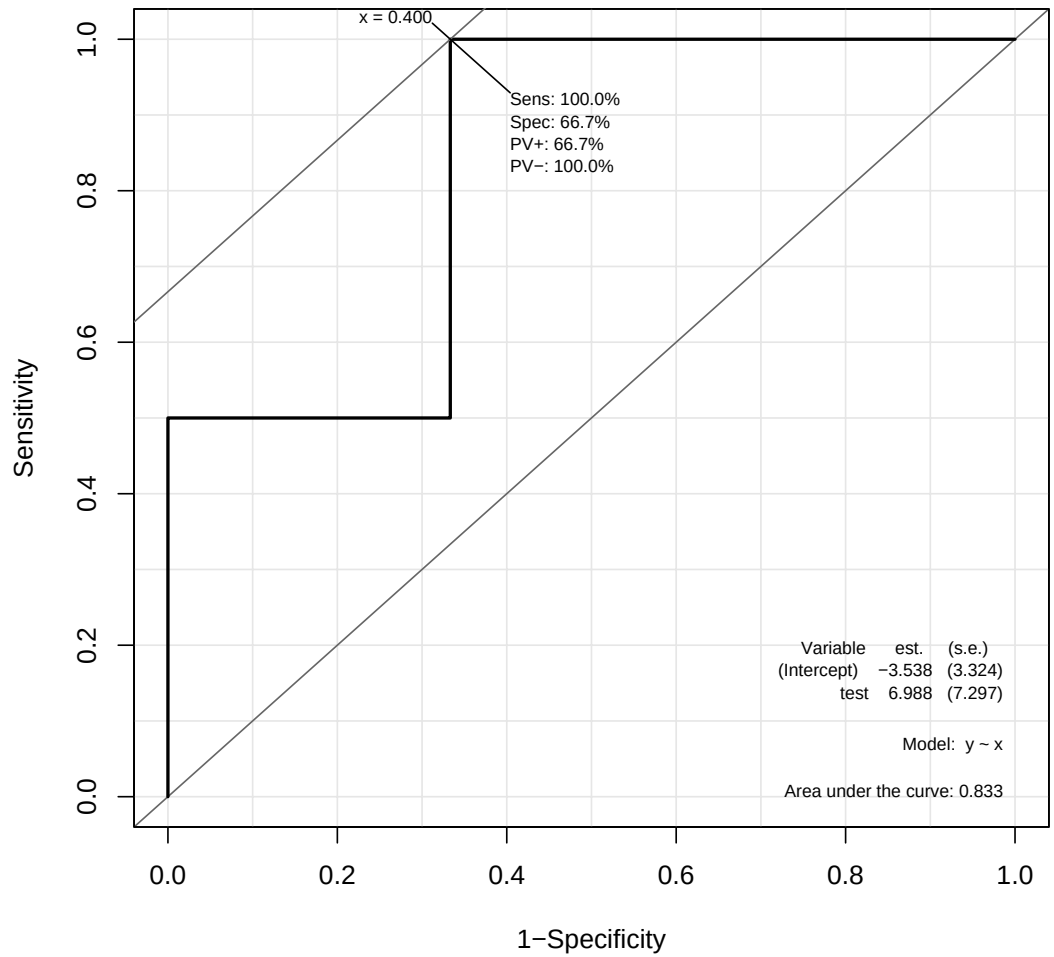
$$f(\hat{Y} = 0|Y = 0) = 1 - f(x \geq c|Y = 0) = 1 - S_0(c)$$

Falsch Positiv = 1- Spezifität:

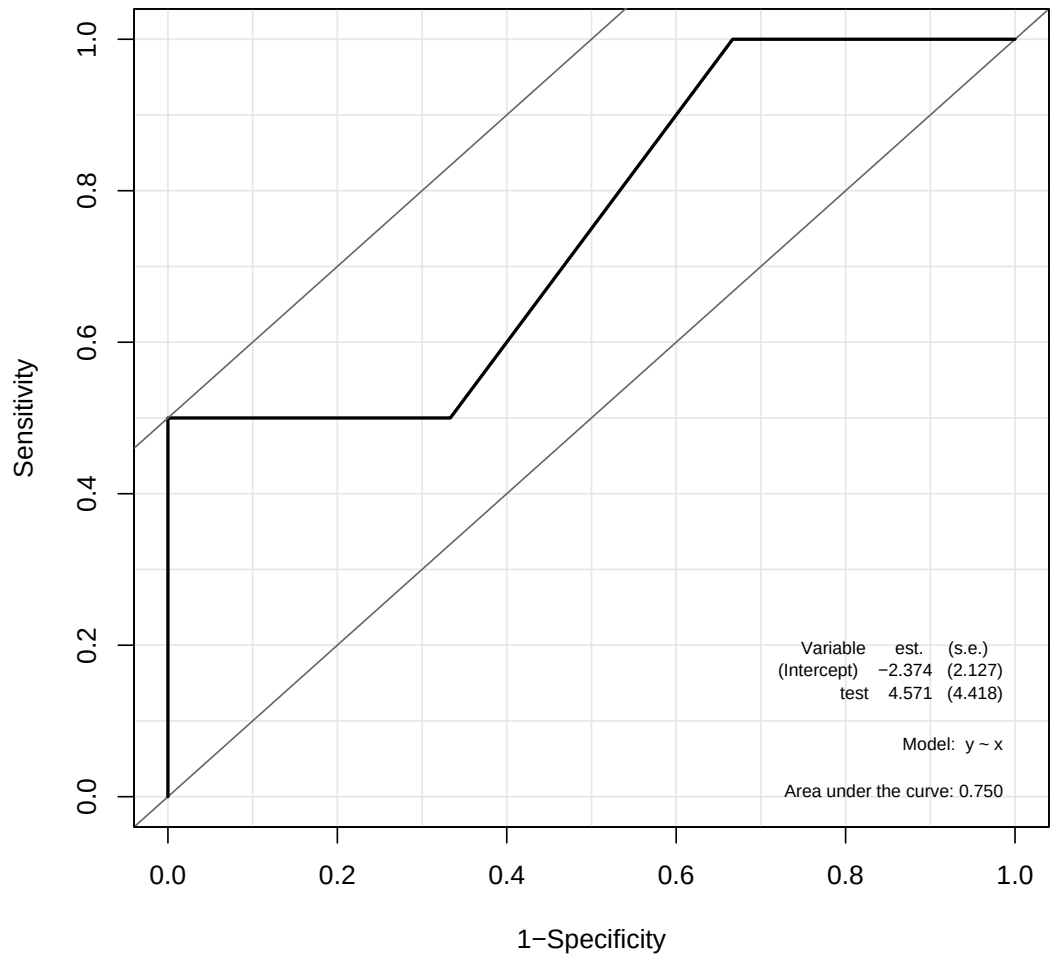
$$f(\hat{Y} = 1|Y = 0) = f(x \geq c|Y = 0) = S_0(c)$$

Die ROC-Kurve besteht aus den Punkten $(S_0(c), S_1(c))$

Beispiel für ROC-Kurve



Beispiel für ROC-Kurve mit Bindung



Maß zur Bewertung der Kurve: AUC

$$AUC = \int_{t=0}^{1} ROC(t)dt \tag{3.7}$$

Dies stellt die Fläche unter der Kurve dar.

Es gilt:

$$AUC = \frac{N_C + 0.5 * N_E}{N} \tag{3.8}$$

Dabei bezeichnet N_C die Anzahl der konkordanten Paare, N_E die Anzahl der identischen Paare, und N die Anzahl der Paare mit unterschiedlichem Y.

GINI- Koeffizient

Normierte Fläche zwischen Winkelhalbierender und ROC- Kurve

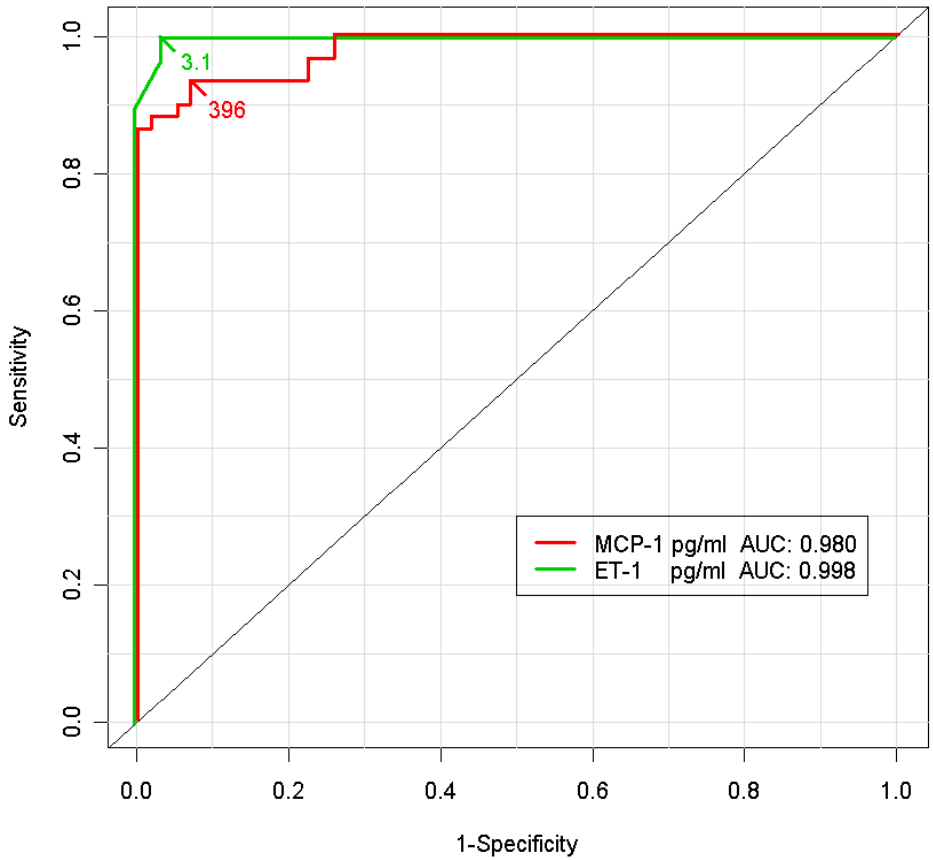
$$GINI = 2 \cdot (AUC - \frac{1}{2}) = 2 \cdot AUC - 1 \tag{3.9}$$

$$GINI = \frac{N_C - N_D}{N} \tag{3.10}$$

N_C: Anzahl der konkordanten Paare
N_D: Anzahl der diskordanten Paare
N: Anzahl der Paare mit ungleichem Y
N = $n_0 \cdot n_1$ mit n_i Anzahl der Daten mit Y=i.

Der GINI entspricht dem Somers D.

Beispiel: Stress induzierter Herzinfarkt



Korrelationsmatrix

Bei mehr als zwei Merkmalen werden die Korrelationen häufig in Form einer Matrix dargestellt.
Auf der Hauptdiagonalen stehen 1er.
Die Matrix ist symmetrisch.

$$\begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix}$$