Bivariate deskriptive Statistik: Regression und Korrelation

Statistik I: Bivariate Statistik, Teil I

1 Bivariate deskriptive Statistik

2 Formen funktionaler Zusammenhänge

3 Streudiagramm

4 Lineare Regression

(Regressionsgerade, Vorhersagefehler, Optimierungs-kriterien, Regressionsgewichte, Vorhersagerichtung, Standardschätzfehler)

5 Kovarianz

Bivariate Statistik: Literatur

Einführende Literatur

📖 Bortz, J. & Schuster, Ch. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Auflage). Berlin: Springer. [Kap. 10.1 & 10.2 sowie 11.1]

📖 Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz. [Kap. 15.1 & 15.2]

Weiterführende Literatur

📖 Bobko, P. (2001). *Correlation and regression. Applications for industrial organizational psychology and management* (2nd ed.). Thousand Oaks: Sage.

📖 Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.).* Mahwah, NJ: LEA.

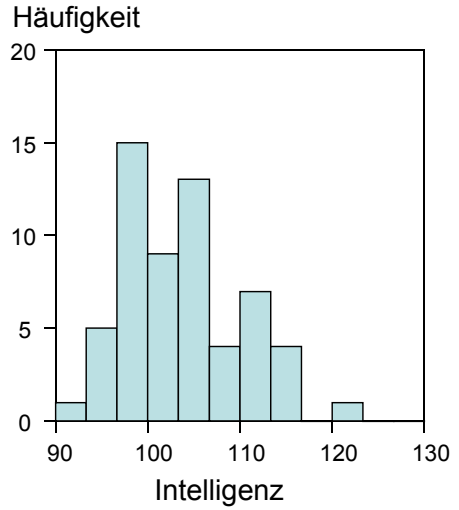
📖 Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

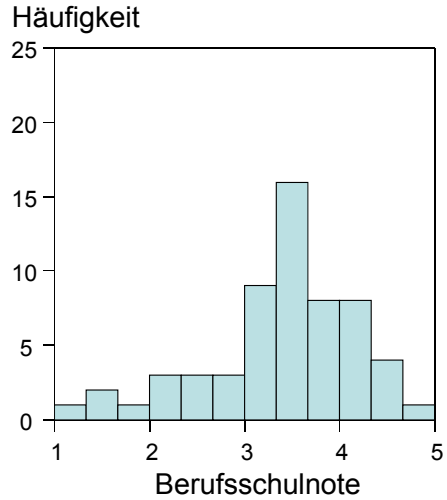
Bivariate deskriptive Statistik

- Bisher haben wir uns Variablen einzeln angeschaut (univariat): Verteilungen sowie Statistiken wie Mittelwert oder Standardabweichung etc.

- In der Psychologie interessiert man sich aber häufig dafür, wie zwei Variablen zusammenhängen oder kovariieren (=bivariate Zusammenhänge), z.B.:
  - Schließen intelligentere Azubis ihre Ausbildung erfolgreicher ab?
  - Wie verändert sich die Laufzeit einer Ratte in einem Labyrinth über die Zeit (Versuchsdurchgänge)?
  - Wie hängt der Aktivierungsgrad mit der Prüfungsleistung zusammen?

- Zudem kann man, wenn man weiß, wie zwei Variablen zusammenhängen, aus der Kenntnis der Ausprägung einer Person in der einen Variablen (mit einer gewissen Genauigkeit) vorhersagen, wie deren Ausprägung in der anderen Variablen ausfallen wird.



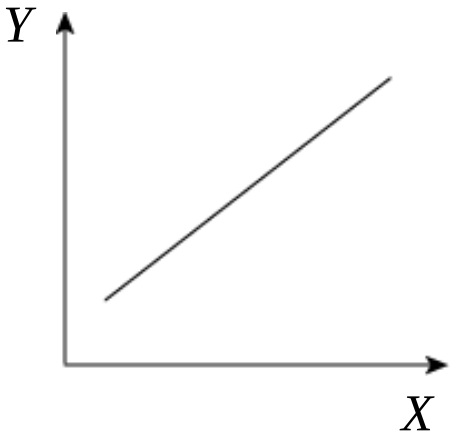


$\bar{x} = 103.6, s = 6.1$ $N = 59$ $\bar{x} = 3.4, s = 0.8$

Formen funktionaler Zusammenhänge

- Wie können nun solche Zusammenhänge überhaupt aussehen?

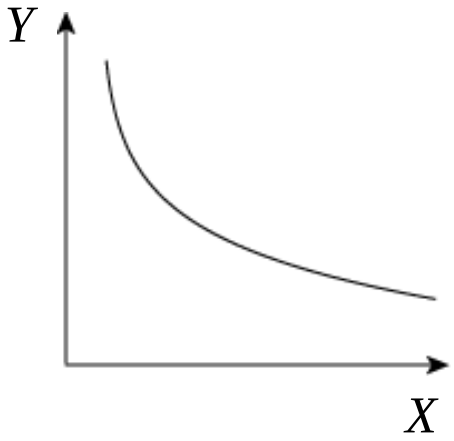
linear



X: Intelligenz

Y: Ausbildungserfolg

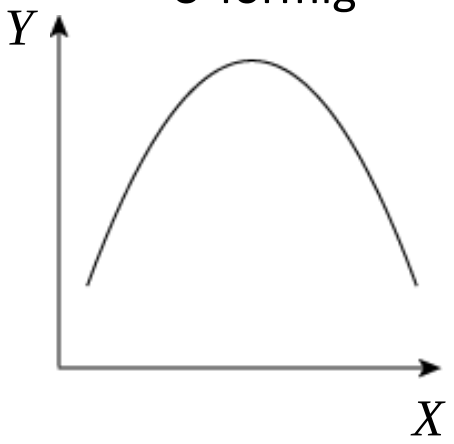
monoton fallend



X: Versuchsdurchgang

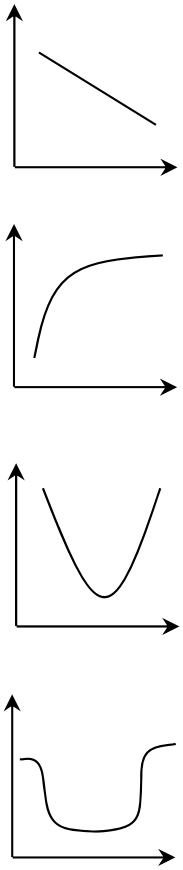
Y: Laufzeit einer Ratte in einem Labyrinth

umgekehrt U-förmig



X: Aktivierung

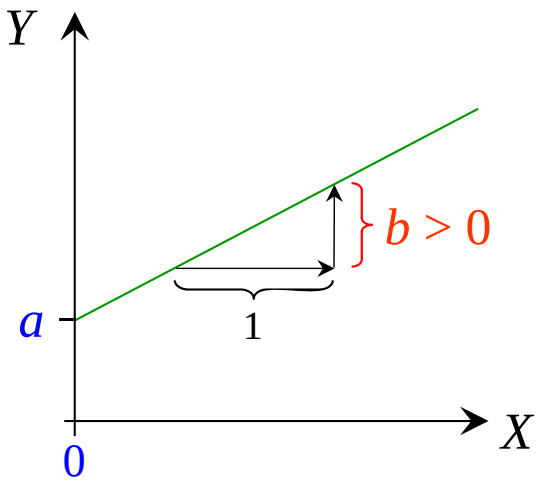
Y: Prüfungsleistung

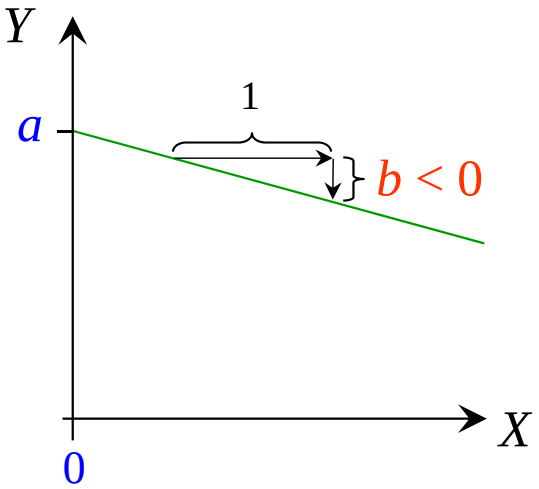


…

Linearer Zusammenhang

- Ein linearer Zusammenhang wird beschrieben durch eine lineare Gleichung:





Kriterium (svariable)

Prädiktor (variable)

$$Y = a + b \cdot X$$

additive Konstante a

multiplikative Konstante b

Y - Achsenab-schnitt (intercept)

Steigung (slope)

Beispiel: Temperatur-Umrechnung

Y = Grad in Fahrenheit (°F)

X = Grad in Celsius (°C)

$Y = 32 + 1.8 \cdot X$

$x = 20°C \Rightarrow y = 32 + 1.8 \cdot 20°C = 68°F$

1 Bivariate deskriptive Statistik

2 Formen funktionaler Zusammenhänge

3 Streudiagramm

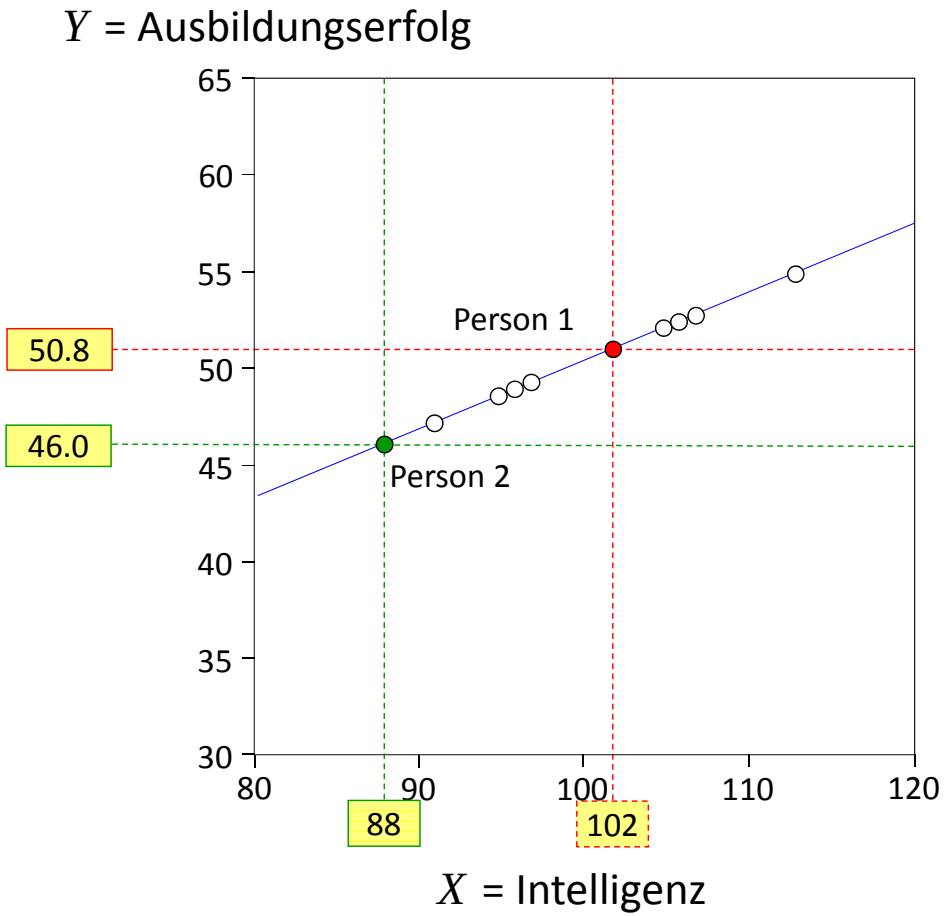
4 Lineare Regression

(Regressionsgerade, Vorhersagefehler, Optimierungs-kriterien, Regressionsgewichte, Vorhersagerichtung, Standardschätzfehler)

5 Kovarianz

Streudiagramm

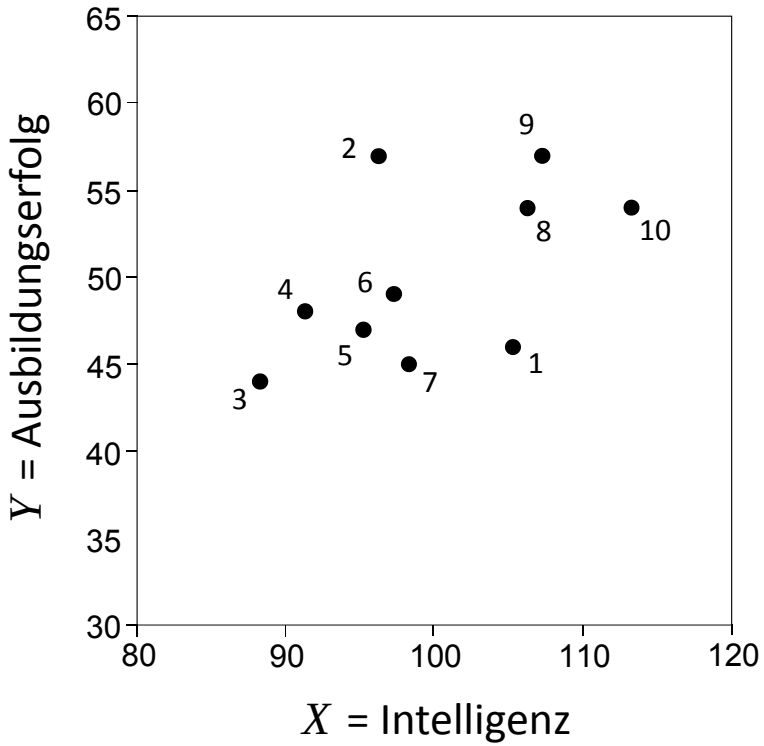
- Wie können wir feststellen, ob zwischen zwei Variablen ein linearer (oder auch anderer) Zusammenhang besteht? Z.B. zwischen Intelligenz und Ausbildungserfolg?
- Zunächst können wir die beiden Variablen an einer Stichprobe von (hier nur $n = 10$) Personen erheben und diese dann in ein Streudiagramm (scatterplot) eintragen.



Person Nr.	X = Intelligenz	Y = Ausbildungserfolg
1	102	50.8
2	88	46.0
3	91	47.1
4	105	52.0
5	95	48.4
6	96	48.8
7	97	49.2
8	106	52.3
9	107	52.7
10	113	54.8
Min =	88	46.0
Max =	113	54.8

Streudiagramm

- Bei empirisch erhobenen Daten ist es aber sehr unwahrscheinlich, dass ein solch perfekter linearer Zusammenhang besteht. Vielmehr muss man damit rechnen, dass die funktionalen Zusammenhänge mit Fehlern überlagert sind (z.B. Messfehlern bei der Erfassung der Intelligenz und des Ausbildungserfolgs).
- Beispiel: Nehmen wir an, es hätten sich für 10 Personen folgende Daten in den beiden Variablen X = Intelligenz und Y = Ausbildungserfolg ergeben:



Vp-Nr.	X	Y
1	105	46
2	96	57
3	88	44
4	91	48
5	95	47
6	97	49
7	98	45
8	106	54
9	107	57
10	113	54

1 Bivariate deskriptive Statistik

2 Formen funktionaler Zusammenhänge

3 Streudiagramm

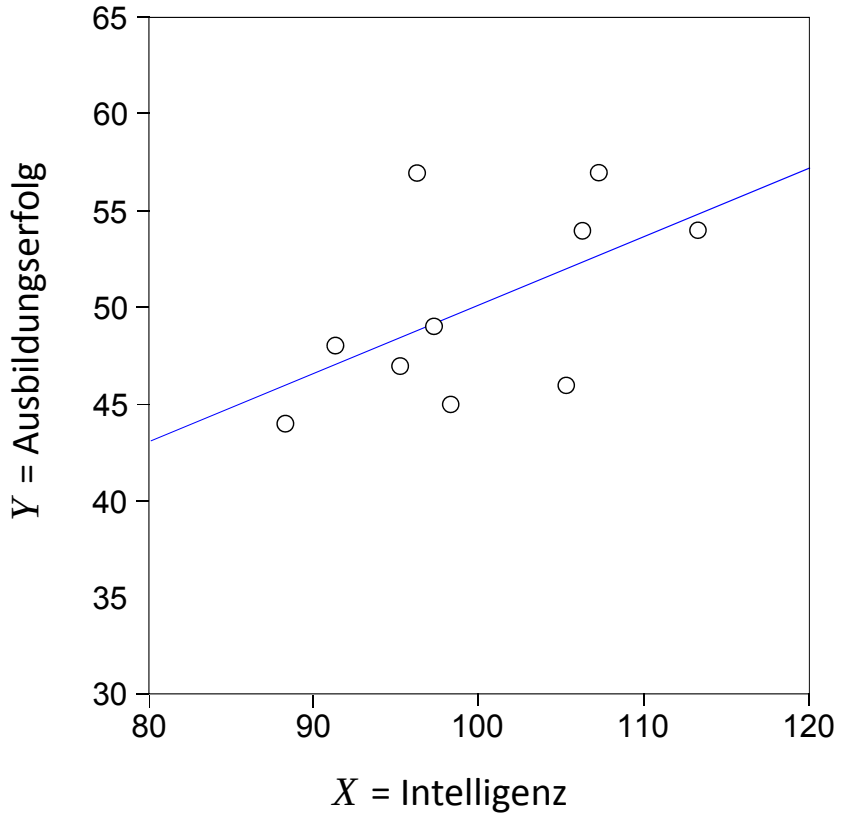
4 Lineare Regression

(Regressionsgerade, Vorhersagefehler, Optimierungs-kriterien, Regressionsgewichte, Vorhersagerichtung, Standardschätzfehler)

5 Kovarianz

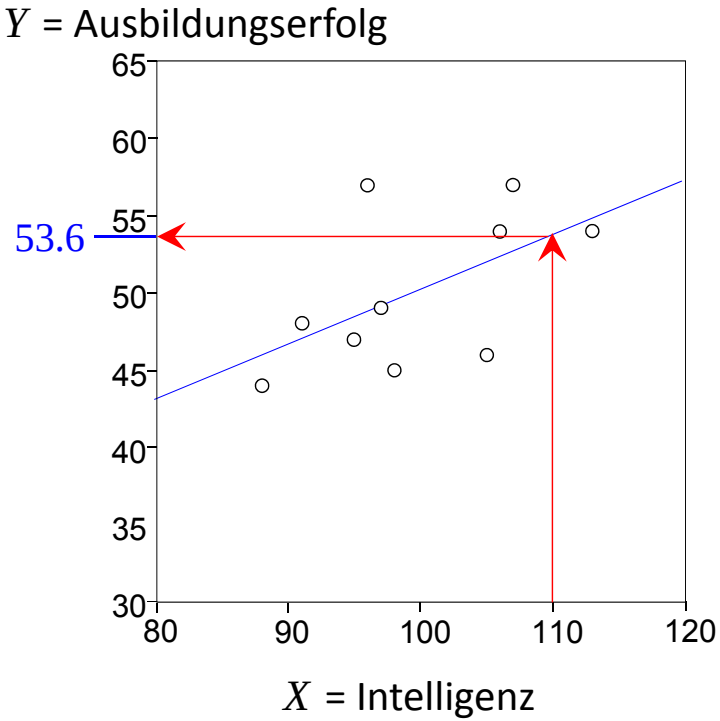
Lineare Regression: Regressionsgerade

- Wir suchen nun eine Gerade, die den linearen Zusammenhang zwischen beiden Variablen „bestmöglich“ charakterisiert. Wir bezeichnen diese Gerade als Regressionsgerade und sprechen von einer „Regression von Y auf X“ (oder „Variable Y wird auf X regrediert“).



Lineare Regression: Regressionsgerade

- Wenn wir eine Regressionsgerade hätten, könnten wir sie zur Vorhersage nutzen:

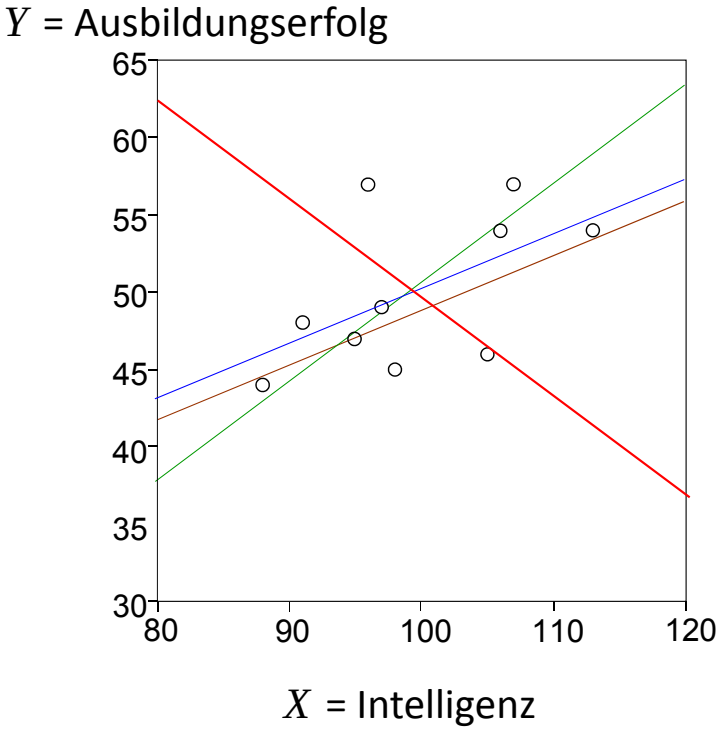


$x = 110$

vorhergesagtes $y = \hat{y} = 53.6 = 15.12 + 0.35 \cdot 110$

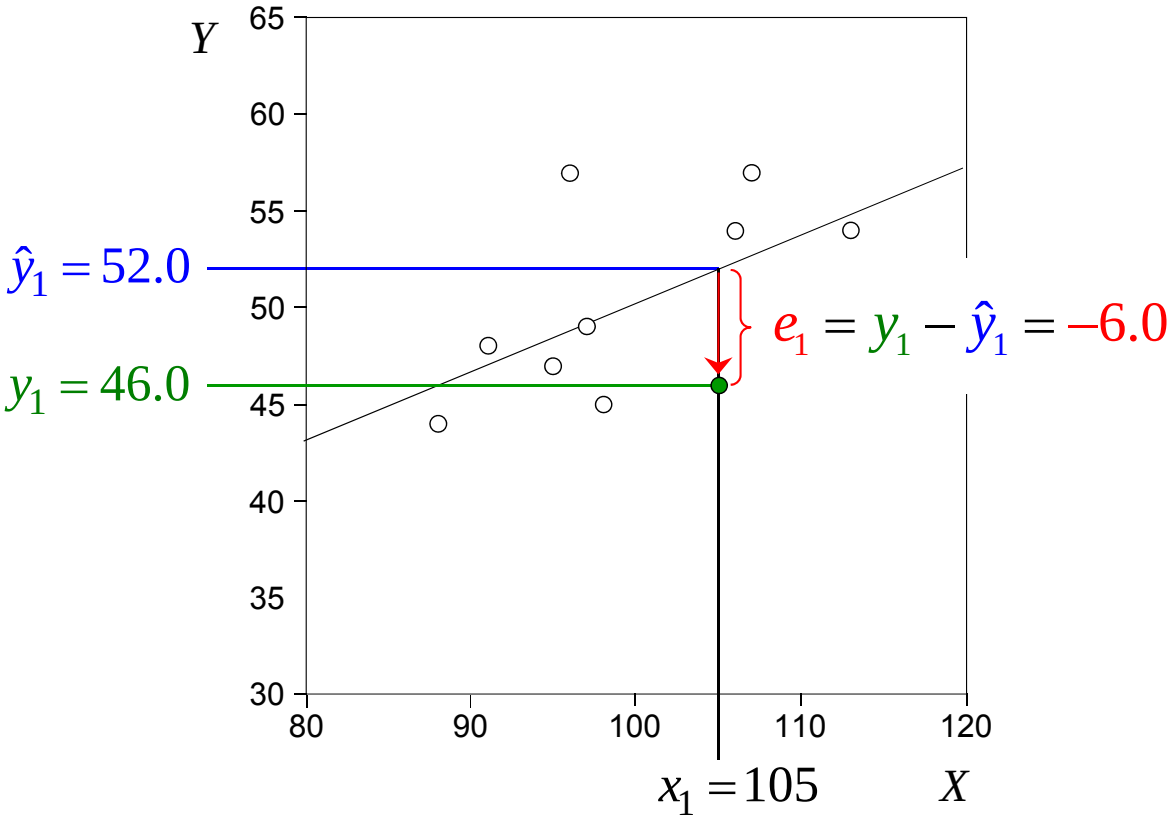
„y-Dach“, „y-hat“

- Aber: Welche ist die „optimale“ Gerade für eine solche Vorhersage, die „nahe“ an den Punkten liegt?



Lineare Regression: Vorhersagefehler

- Um zu einer „optimalen“ Gerade zu kommen betrachten wir den Vorhersagefehler, den wir bei einer beliebig gewählten Gerade machen (hier z.B. mit $a = 15.12$ und $b = 0.35$).



x_i … beobachteter Wert der Person i im Prädiktor X

y_i … beobachteter Wert der Person i im Kriterium Y

$\hat{y}_i$ … vorhergesagter Wert der Person i im Kriterium Y

e_i … Fehler(wert) (=Residuum) in der Vorhersage von Person i

$1 \leq i \leq n$ ($n = 10$ … Zahl der Personen)

- Die Regressionsgleichung lautet:

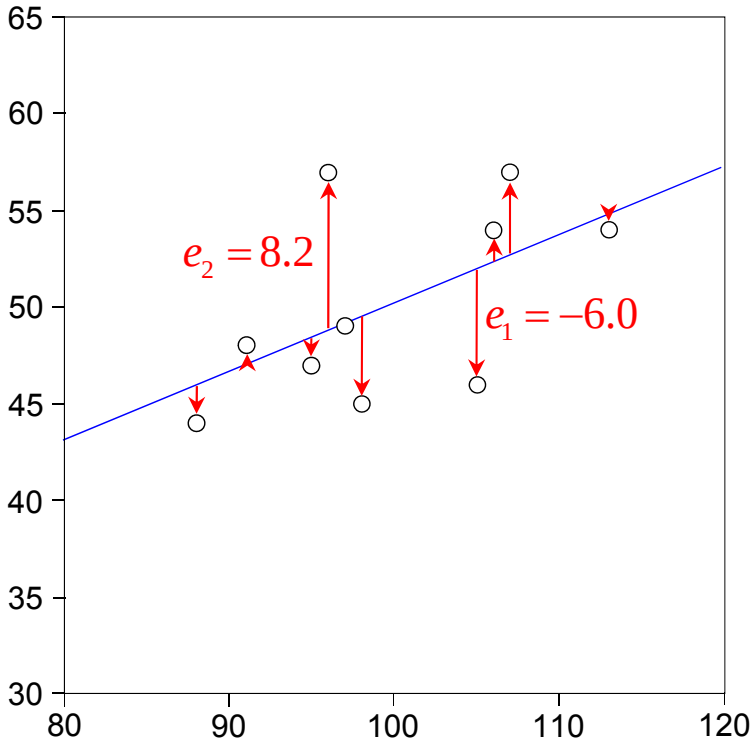
$$\hat{y}_i = a + b \cdot x_i \qquad = 15.12 + 0.35 \cdot x_i$$

$$\Leftrightarrow \; y_i = a + b \cdot x_i + e_i = 15.12 + 0.35 \cdot x_i + e_i$$

Lineare Regression: Optimierungskriterien

- Wie könnte man die Gerade so legen, dass sie möglichst nahe an allen Punkten liegt?

Y = Ausbildungserfolg



X = Intelligenz

Wähle die Gerade so, dass …

- Vorschlag 1: … die Summe aller Residuen e_i möglichst klein ist

  also: $\sum_{i=1}^{n} e_i = -6 + 8.2 + \ldots \rightarrow \min$ ☹

- Vorschlag 2: … die Summe der Beträge aller e_i (=Abstände von der Regressionsgeraden) möglichst klein ist

  also: $\sum_{i=1}^{n} |e_i| = |-6| + |8.2| + \ldots \rightarrow \min$ 😐

- Vorschlag 3: … die Summe der quadrierten e_i möglichst klein ist

  also: $\sum_{i=1}^{n} e_i^2 = (-6)^2 + 8.2^2 + \ldots \rightarrow \min$ ☺

Lineare Regression: Optimierungskriterien

- Das Optimierungskriterium

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

bei der Wahl der Regressionsgeraden, nämlich die Summe der quadrierten Residuen zu minimieren, …

- bezeichnet man auch als Methode der Kleinsten Quadrate (least squares method).
- führt zu einer eindeutigen Lösung, d.h. es gibt genau eine Gerade, die dieses Kriterium erfüllt.
- führt zu einer mathematisch einfach zu handhabenden Lösung (im Gegensatz beispielsweise zur Optimierung von Summen von Beträgen).
- hat die Eigenschaft, dass große Abweichungen durch die Quadrierung besonders stark in die Summe eingehen und daher durch die Wahl der Lage der Regressionsgeraden möglichst vermieden werden.
- führt dazu, dass die Summe der Residuen Null wird (Man beachte die Analogie zu $\overline{x}$):

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

Lineare Regression: Bestimmung der Regressionsgewichte

- Wie kann man nun die hinsichtlich des Kleinste-Quadrate Kriteriums optimalen a und b bestimmen?

$$\hat{y}_i = a + b \cdot x_i$$

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

$$F = \sum_{i=1}^{n} (y_i - [a + b \cdot x_i])^2 = \min$$

(F nach a bzw. b partiell ableiten, erste Ableitungen Null setzen und prüfen ob sie positiv sind.)

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

$$a = \bar{y} - b \cdot \bar{x}$$

vgl. etwa Bortz und Schuster (2010, S. 187)

Lineare Regression: Bestimmung der Regressionsgewichte

- Bestimmung der Regressionsgleichung für Beispieldatensatz mit $n = 10$:

Vp-Nr.	x_i	y_i	x_i^2	$x_i \cdot y_i$	$\hat{y}_i$	e_i	e_i^2
1	105	46	11025	4830	52.00	-6.00	35.96
2	96	57	9216	5472	48.84	8.16	66.66
3	88	44	7744	3872	46.03	-2.03	4.11
4	91	48	8281	4368	47.08	0.92	0.85
5	95	47	9025	4465	48.48	-1.48	2.20
6	97	49	9409	4753	49.19	-0.19	0.03
7	98	45	9604	4410	49.54	-4.54	20.59
8	106	54	11236	5724	52.35	1.65	2.73
9	107	57	11449	6099	52.70	4.30	18.50
10	113	54	12769	6102	54.81	-0.81	0.65
Σ:	996	501	99758	50095	501	0	152.28

→ min

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

$$= \frac{10 \cdot 50095 - 996 \cdot 501}{10 \cdot 99758 - 996^2}$$

$$= 0.3512$$

$$a = \bar{y} - b \cdot \bar{x}$$

$$= \frac{501}{10} - 0.3512 \cdot \frac{996}{10}$$

$$= 15.12$$

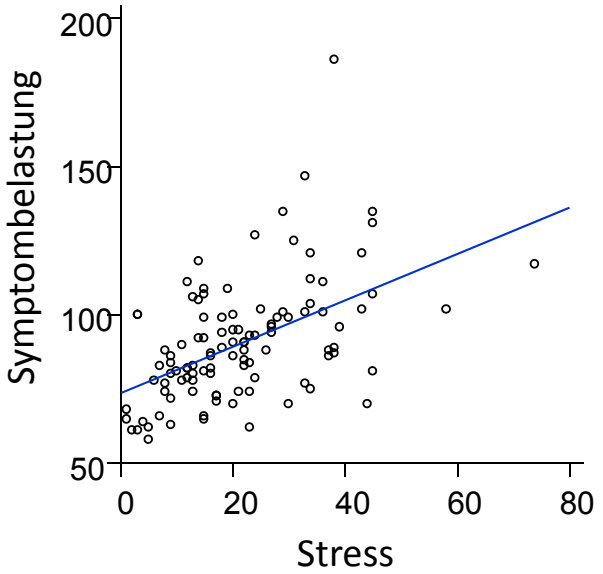
$$\hat{y}_i = a + b \cdot x_i = 15.12 + 0.35 \cdot x_i$$

$$e_i = y_i - \hat{y}_i$$

Lineare Regression: Positiver linearer Zusammenhang

- Besteht ein positiver linearer Zusammenhang, so ...
  - gehen hohe Werte in X eher mit hohen Werten in Y einher.
  - gehen niedrige Werte in X eher mit niedrigen Werten in Y einher.
  - gehen mittlere Werte in X mit mittleren Werten in Y einher.
  - ist das multiplikative Regressionsgewicht positiv $b > 0$.

Wagner, Compass & Howell (1989), $n = 107$

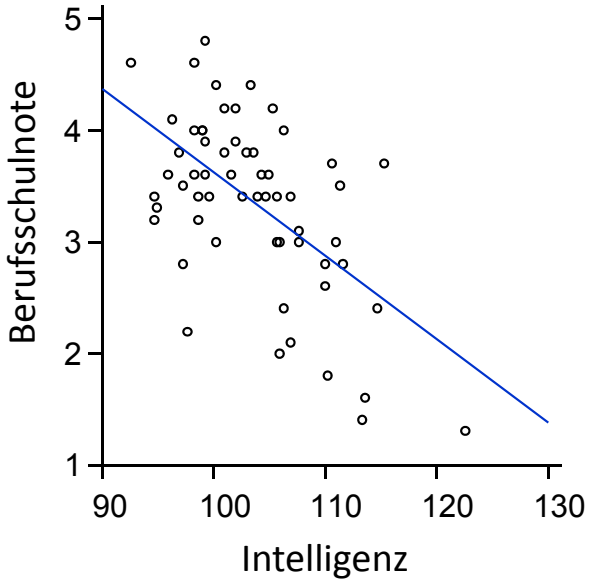


$$\hat{y}_i = 73.89 + 0.78 \cdot x_i$$

Lineare Regression: Negativer linearer Zusammenhang

- Besteht ein negativer linearer Zusammenhang, so ...
  - gehen hohe Werte in X eher mit niedrigen Werten in Y einher.
  - gehen niedrige Werte in X eher mit hohen Werten in Y einher.
  - gehen mittlere Werte in X mit mittleren Werten in Y einher.
  - ist das multiplikative Regressionsgewicht negativ $b < 0$.

Lang (1990),
$n = 59$

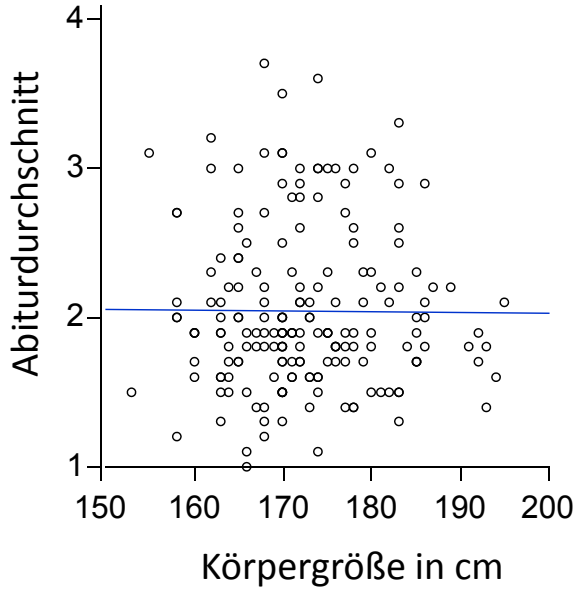


$$\hat{y}_i = 11.08 - 0.08 \cdot x_i$$

Lineare Regression: Kein linearer Zusammenhang

- Besteht kein linearer Zusammenhang, so ...
  - gehen hohe Werte in X gleichermaßen mit hohen wie niedrigen Werten in Y einher.
  - gehen niedrige Werte in X gleichermaßen mit niedrigen wie hohen Werten in Y einher.
  - gehen mittlere Werte in X gleichermaßen mit niedrigen wie hohen Werten in Y einher.
  - ist das multiplikative Regressionsgewicht Null $b = 0$.
- In diesem Fall lautet die Regressionsgleichung nur noch $\hat{y}_i = a$ wobei a dem Mittelwert in Y entspricht. Für alle Personen wird also, gleichgültig welchen Wert sie in X haben, immer der Wert $\bar{y}$ vorhergesagt.
- Der Mittelwert $\bar{y}$ ist auch die beste Vorhersage im Sinne der kleinsten Quadrate, wenn man X gar nicht kennt und Y vorhersagen soll.

Übungsdatensatz,
$n = 187$



$$\hat{y}_i = 2.156 - 0.001 \cdot x_i$$

$$\bar{y} = 2.04$$

Beispiel mit sehr geringem linearen Zusammenhang

Regression in beide Richtungen

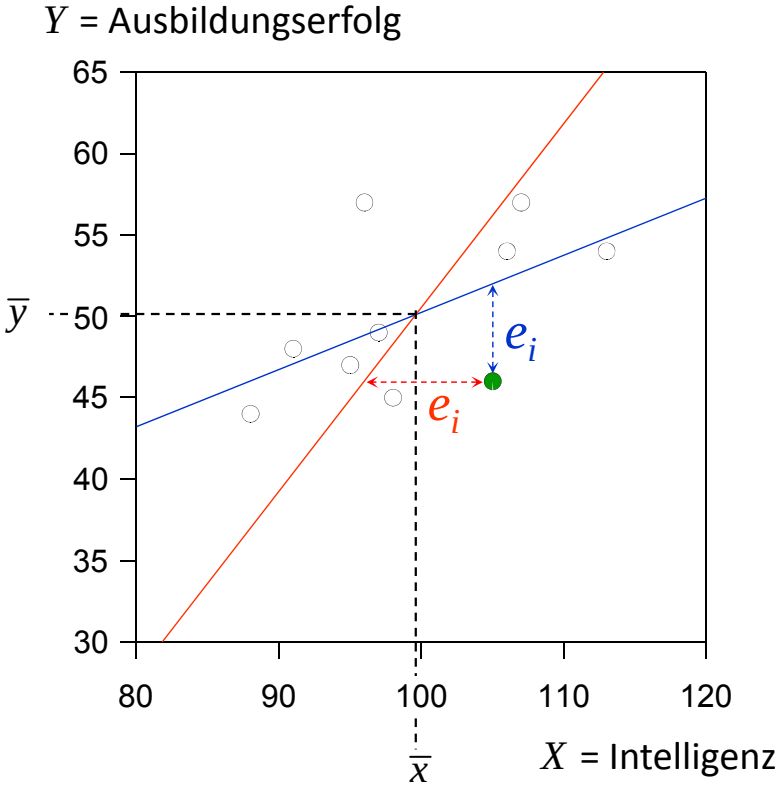
- Kehrt man die Richtung der Fragestellung um und fragt sich, wie gut sich die Variable X aus der Variablen Y vorhersagen lässt, so kann dafür analog eine Regressionsgleichung bestimmt werden. Für den Beispieldatensatz gelten die beiden Gleichungen:

$$y_i = a_{YX} + b_{YX} \cdot x_i + e_i = 15.12 + 0.35 \cdot x_i + e_i$$

$$x_i = a_{XY} + b_{XY} \cdot y_i + e_i = 55.18 + 0.88 \cdot y_i + e_i$$

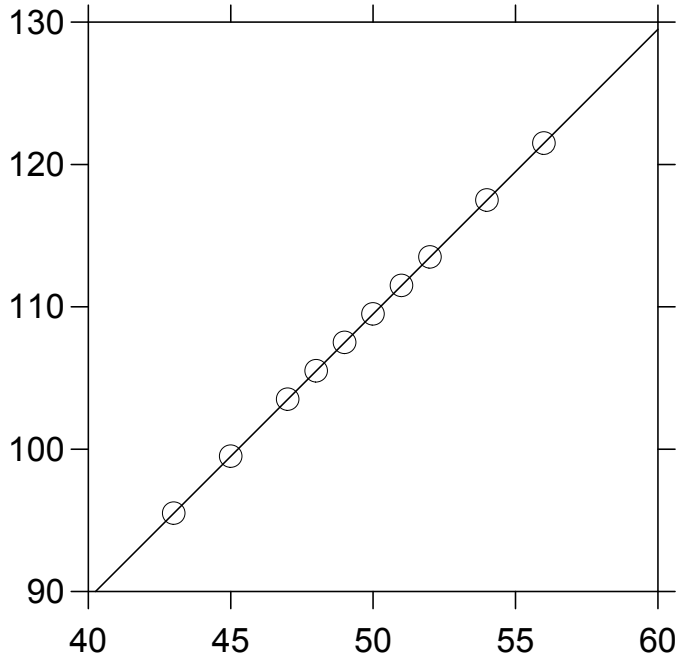
Durch die Indizierung der Regressionsgewichte mit XY bzw. YX wird im Zweifelsfall die Richtung der Vorhersage deutlich.

- Allgemein gilt für die beiden Regressionsgeraden:
  - Sie sind nur bei einem perfekten Zusammenhang identisch, andernfalls unterscheiden sie sich. Dies hängt damit zusammen, dass eine Regressionsgerade die quadratischen Abweichungen in X- und die andere in Y-Richtung minimiert.
  - Sie schneiden sich im Punkt $(\overline{x}, \overline{y})$.

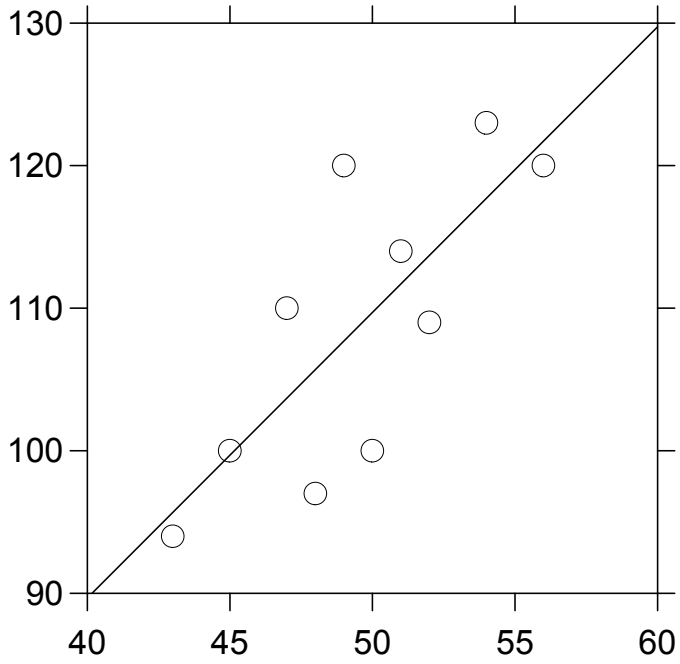


Bewertung der Vorhersage

- Ein Regressionsgleichung/gerade lässt sich für empirische Daten immer bestimmen.
- Als wichtige Information will man daher wissen, wie gut die Vorhersage ist bzw. wie eng der lineare Zusammenhang ist. Aus den Regressionsgewichten lässt sich dies nicht entnehmen (s.u.).



$$\hat{y}_i = 9.5 + 2.0 \cdot x_i$$



$$\hat{y}_i = 9.5 + 2.0 \cdot x_i$$

Standardschätzfehler

- Eine Möglichkeit zur Bestimmung der Vorhersagegüte besteht darin, das Optimierungs-kriterium heranzuziehen. Normiert man dieses an der Stichprobengröße, so erhält man die Varianz der Residuen (=Fehlervarianz=Residualvarianz):

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

$$\sum_{i=1}^{n}(e_i - \bar{e})^2 = \sum_{i=1}^{n}(e_i - 0)^2 = \sum_{i=1}^{n} e_i^2$$

Warum durch $n-2$ dividiert wird, werden wir später sehen.

- Gebräuchlicher ist die Verwendung der Wurzel aus dieser Varianz, die man als Standard-schätzfehler s_E (standard error of estimate) bezeichnet:

$$s_E = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

- Bei einer perfekten Vorhersage ist der Standardschätzfehler 0. Je stärker die Punkte um die Regressionsgerade streuen, je schlechter also die Vorhersage, desto größer wird s_E.

Standardschätzfehler

- Bestimmung des Standardschätzfehlers für Beispieldatensatz mit $n = 10$:

Vp-Nr.	x_i	y_i	x_i^2	$x_i \cdot y_i$	$\hat{y}_i$	e_i	e_i^2
1	105	46	11025	4830	52.00	-6.00	35.96
2	96	57	9216	5472	48.84	8.16	66.66
3	88	44	7744	3872	46.03	-2.03	4.11
4	91	48	8281	4368	47.08	0.92	0.85
5	95	47	9025	4465	48.48	-1.48	2.20
6	97	49	9409	4753	49.19	-0.19	0.03
7	98	45	9604	4410	49.54	-4.54	20.59
8	106	54	11236	5724	52.35	1.65	2.73
9	107	57	11449	6099	52.70	4.30	18.50
10	113	54	12769	6102	54.81	-0.81	0.65
Σ:	996	501	99758	50095	501	0	152.28

$$e_i^2 = (y_i - \hat{y}_i)^2$$

$$s_E = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

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

$$= \sqrt{\frac{152.28}{10-2}}$$

$$= 4.36$$

1 Bivariate deskriptive Statistik

2 Formen funktionaler Zusammenhänge

3 Streudiagramm

4 Lineare Regression

(Regressionsgerade, Vorhersagefehler, Optimierungs-kriterien, Regressionsgewichte, Vorhersagerichtung, Standardschätzfehler)

5 Kovarianz

Kovarianz

- Eine andere Möglichkeit, die Enge des Zusammenhangs zu quantifizieren besteht darin, die Kovarianz *cov* zu bestimmen.

$$cov = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} a(x_i) \cdot a(y_i)}{n-1}$$

$a(x_i)$ = Abweichungswert $x_i - \bar{x}$

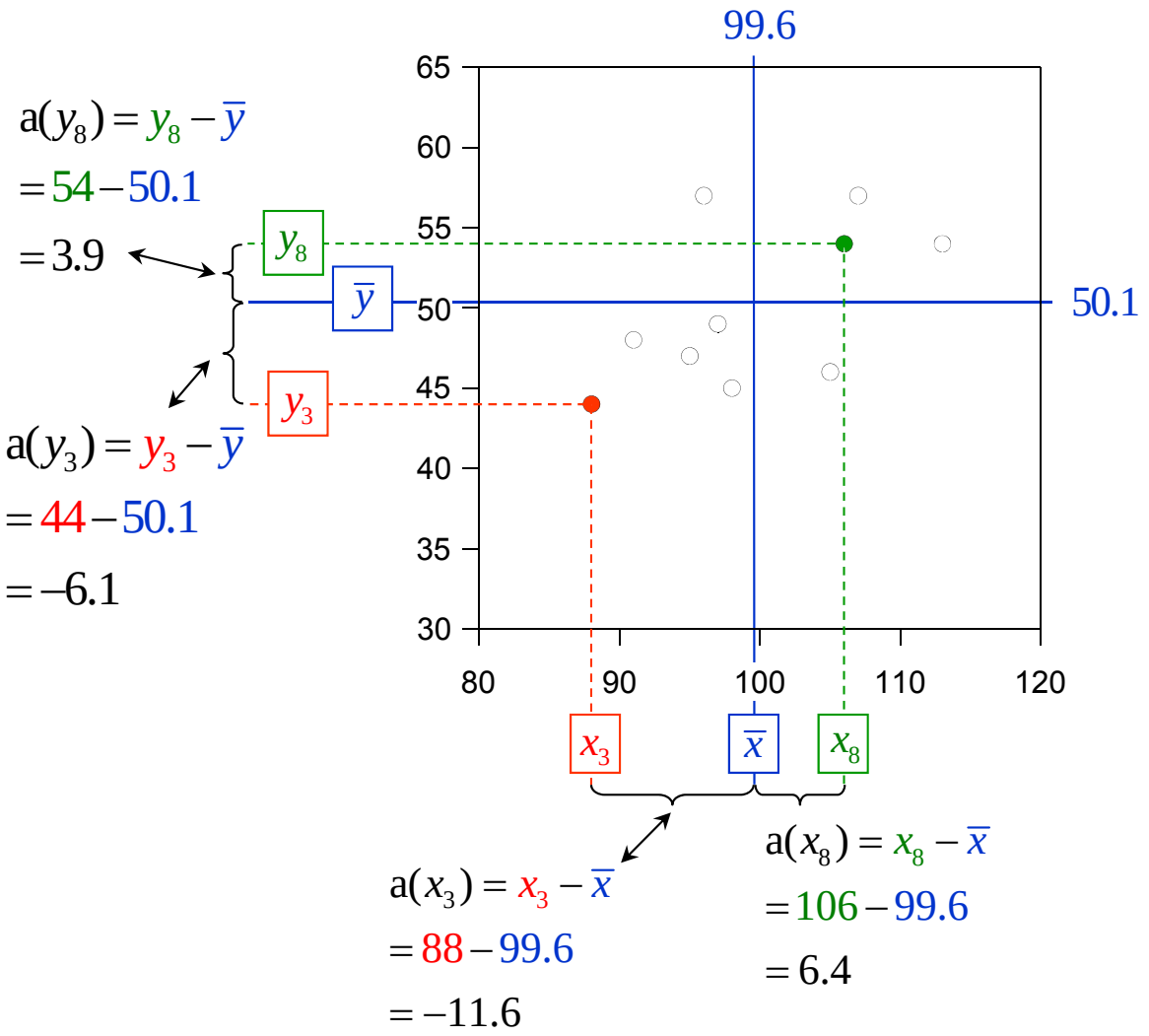
$a(y_i)$ = Abweichungswert $y_i - \bar{y}$

- Die Kovarianz zwischen zwei Variablen wird umso (positiv) größer, je mehr positive Abweichungen vom Mittelwert in X mit positiven Abweichungen vom Mittelwert in Y sowie negative Abweichungen in X mit negativen Abweichungen in Y einhergehen. Die Kreuzprodukte $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ sind dann positiv.
- Die Kovarianz wird umso (negativ) kleiner, je mehr positive Abweichungen in X mit negativen Abweichungen in Y sowie negative Abweichungen in X mit positiven Abweichungen in Y einhergehen. Die Kreuzprodukte sind dann negativ.
- Besteht kein Zusammenhang und treten positive Abweichungen in X gleichermaßen mit positiven wie negativen in Y und gilt für die negativen Abweichungen in X das gleiche, so heben sich positive und negative Kreuzprodukte auf und es resultiert eine Kovarianz von 0.

Kovarianz

Kovarianz bei einem positiven Zusammenhang wie im Beispieldatensatz:

Vp	x_i	y_i	$a(x_i)$	$a(y_i)$	$a(x_i)\cdot a(y_i)$
1	105	46	5.4	-4.1	-22.14
2	96	57	-3.6	6.9	-24.84
3	88	44	-11.6	-6.1	70.76
4	91	48	-8.6	-2.1	18.06
5	95	47	-4.6	-3.1	14.26
6	97	49	-2.6	-1.1	2.86
7	98	45	-1.6	-5.1	8.16
8	106	54	6.4	3.9	24.96
9	107	57	7.4	6.9	51.06
10	113	54	13.4	3.9	52.26



$$cov = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) = \frac{1}{n-1}\cdot\sum_{i=1}^{n}a(x_i)\cdot a(y_i)$$

$$= \frac{1}{10-1}\cdot\left[5.4\cdot(-4.1)+(-3.6)\cdot 6.9+\ldots+13.4\cdot 3.9\right] = 21.71$$

Kovarianz

- Die Kovarianz weist folgende Eigenschaften auf:
  - Sie zeigt die Enge und Richtung des linearen Zusammenhangs zwischen zwei intervallskalierten Variablen X und Y an. Ist $cov = 0$, so besteht kein linearer Zusammenhang.
  - Das Vorzeichen der Kovarianz gibt an, ob es sich um einen positiven oder negativen linearen Zusammenhang handelt.
  - Die Kovarianz ist invariant gegenüber Additionen von Konstanten zu X oder Y: $cov(X + a, Y + b) = cov(X, Y)$.
  - Bei einer Multiplikation von X oder Y mit einer positiven Konstanten wird auch die Kovarianz um diese Faktoren größer: $cov(c \cdot X, d \cdot Y) = c \cdot d \cdot cov(X, Y)$ mit $c, d > 0$.
  - Die Kovarianz ist also maßstabsabhängig.
  - Weist X oder weist Y keine Varianz auf, so ist die Kovarianz 0.

- Man kann die Kovarianz alternativ auch nach folgender Formel berechnen (die besser zur Programmierung geeignet ist):

$$cov = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n - 1}$$

Statistik I: Bivariate Statistik, Teil II

6 Produkt-Moment Korrelation

(Korrelation, Quadratsummen, Determinations-koeffizient, verzerrende Einflüsse)

7 Partialkorrelation

8 Berechnung der Statistiken mittels SPSS

9 Weitere Korrelationskoeffizienten

(Spearmans r_s, Kendalls τ ...)

10 Deutung von Korrelationen

Produkt-Moment Korrelation

- Um ein Zusammenhangsmaß zu erhalten, dass unabhängig von den Maßstäben der beiden Variablen X und Y ist, normiert man die Kovarianz an den Standardabweichungen der Variablen X (s_X) und Y (s_Y) und erhält damit als Statistik die Produkt-Moment Korrelation r (nach Pearson):

$$r = \frac{cov}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_X \cdot s_Y} \quad \text{für } s_X \neq 0 \text{ und } s_Y \neq 0$$

Karl Pearson
(1857-1936)

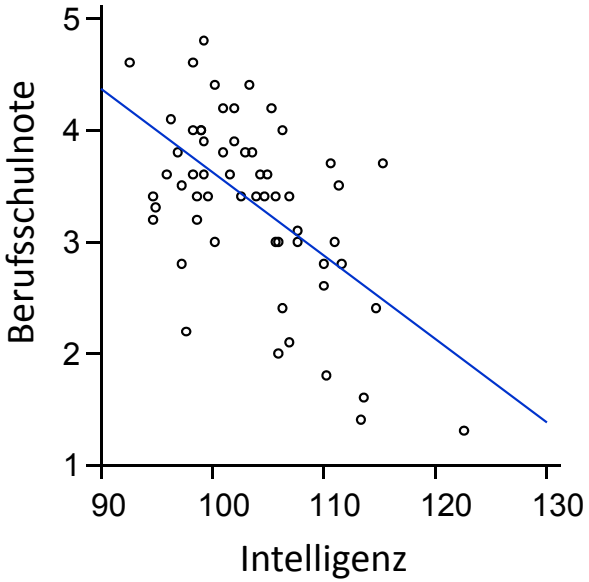
- Durch die Normierung liegt r immer im Wertebereich zwischen -1 und 1. Es gilt:
  - Bei einem perfekten positiven, linearen Zusammenhang ist $r = 1$.
  - Bei einem perfekten negativen, linearen Zusammenhang ist $r = -1$.
  - Besteht kein linearer Zusammenhang, so ist $r = 0$.

- Beispieldatensatz mit $n = 10$, $cov = 21.711$, $s_X = 7.863$ und $s_Y = 4.954$:

$$r = \frac{cov}{s_X \cdot s_Y} = \frac{21.711}{7.863 \cdot 4.954} = 0.56$$

Produkt-Moment Korrelation

Lang (1990),
$N = 59$

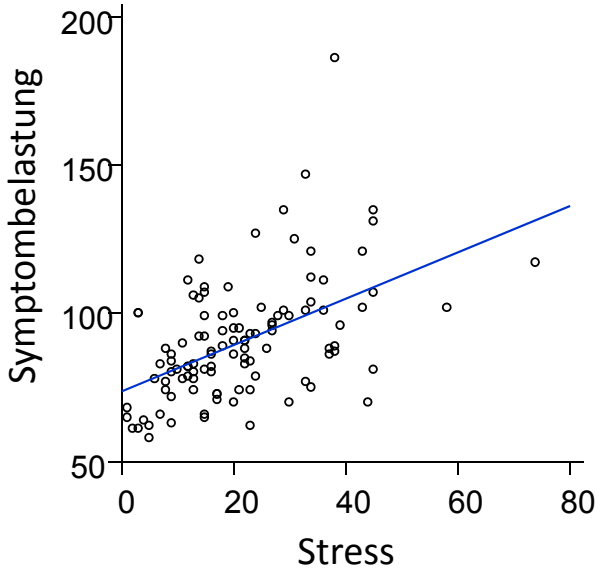


$$\hat{y}_i = 11.08 - 0.08 \cdot x_i$$

negativer linearer Zusammenhang

$r = -0.58$

Wagner, Compass & Howell (1989), $N = 107$

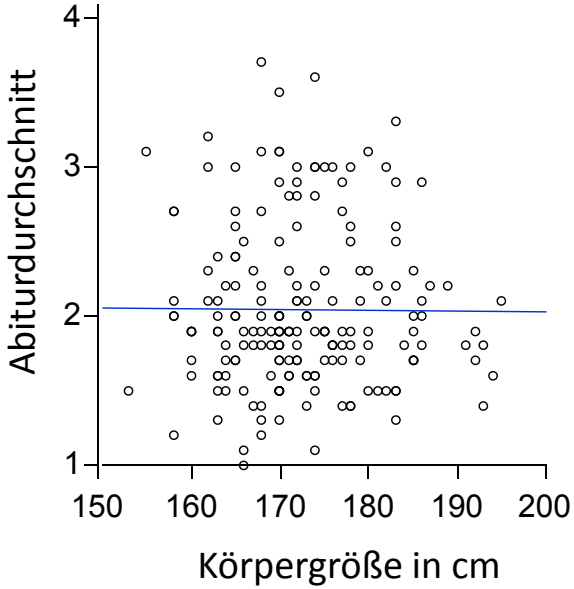


$$\hat{y}_i = 73.89 + 0.78 \cdot x_i$$

positiver linearer Zusammenhang

$r = 0.51$

Übungsdatensatz,
$N = 187$



$$\hat{y}_i = 2.16 - 0.00 \cdot x_i$$

kein linearer Zusammenhang

$r = -0.01$

Produkt-Moment Korrelation

- Schreibt man die Formel wie folgt etwas anders

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_X \cdot s_Y} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \underbrace{\frac{x_i - \bar{x}}{s_X}}_{z(x_i)} \cdot \underbrace{\frac{y_i - \bar{y}}{s_Y}}_{z(y_i)}$$

so erkennt man, dass die Produkt-Moment Korrelation auch der „mittleren“ Summe der Produkte der korrespondierenden *z*-Werte beider Variablen entspricht:

$$r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} z(x_i) \cdot z(y_i)$$

- Besser zur Programmierung ist die folgende Formel geeignet:

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

Produkt-Moment Korrelation

- Die Produkt-Moment Korrelation r weist folgende Eigenschaften auf:
  - Sie zeigt die Enge und Richtung des linearen Zusammenhangs zwischen zwei intervallskalierten Variablen X und Y an.
  - r ändert sich nicht bei ("ist invariant gegenüber") linearen Transformationen, d.h. die Multiplikation der Werte in X oder Y mit einer beliebigen positiven Konstanten und/oder die Addition einer beliebigen Konstante zu den Werten in X und Y ändert r nicht: $r(a + b \cdot X, c + d \cdot Y) = r(X, Y)$ mit $b, d > 0$.
  - Das bedeutet: r ist unabhängig vom Maßstab der Variablen.
  - r hängt mit dem multiplikativen Regressionsgewicht wie folgt zusammen:

    $$r = b_{YX} \cdot \frac{s_X}{s_Y} = b_{XY} \cdot \frac{s_Y}{s_X}$$

    r weist damit immer das gleiche Vorzeichen wie b auf.

Produkt-Moment Korrelation

- Weitere Eigenschaften der Produkt-Moment Korrelation r:
  - Liegen X und Y z-standardisiert vor (d.h. die Standardabweichungen in X und Y sind 1), dann entspricht die Korrelation der Steigung der Regressionsgeraden b. Zudem ist dann der Achsenabschnitt $a = 0$, d.h. die Regressionsgerade geht durch den Ursprung des Koordinatensystems. Es gilt also:

$$\hat{z}(y_i) = a + b \cdot \mathrm{z}(x_i) = r \cdot \mathrm{z}(x_i)$$

  - r hängt wie folgt mit dem Standardschätzfehler zusammen:

$$s_E = s_Y \cdot \sqrt{(1 - r^2) \cdot \frac{n-1}{n-2}}$$

Untenstehende in vielen Statistiklehrbücher angegebene vereinfachte Näherung unterscheidet sich davon nur bei großem n kaum:

$$s_E = s_Y \cdot \sqrt{1 - r^2}$$

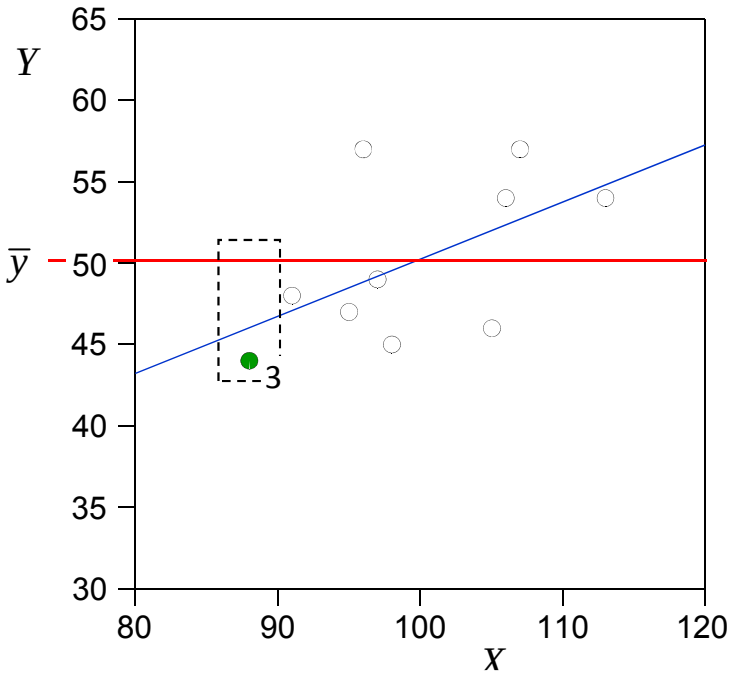
Produkt-Moment Korrelation

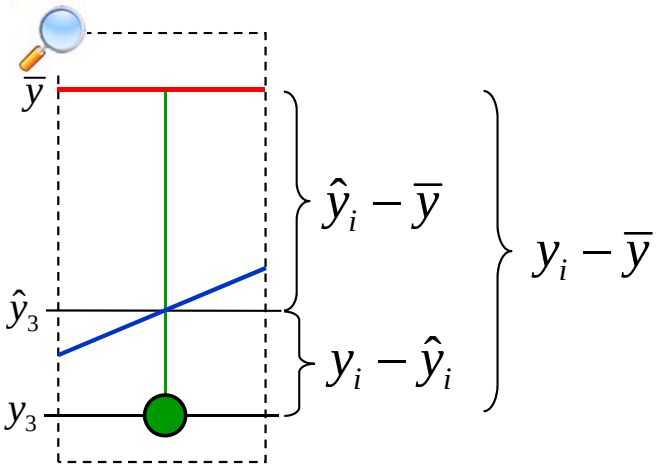
- Betrachtet man die Beziehungen zwischen den Abweichungen der beobachteten und der vorhergesagten y-Werte und dem Mittelwert in Y so gilt:

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

- Ohne Kenntnis der Variablen X würde man für eine Person i den Wert $\bar{y}$ vorhersagen und damit den Fehler $y_i-\bar{y}$ machen. Bei Kenntnis des Wertes x_i und Einsatz der Regressionsgleichung machen wir den Fehler $e_i=y_i-\hat{y}_i$.

- In dem Maße, in dem die $\hat{y}_i$-Werte von $\bar{y}$ abweichen verringert sich der Fehler $y_i-\hat{y}_i$ relativ zu $y_i-\bar{y}$.





Produkt-Moment Korrelation

- Man kann nun zeigen, dass die additive Beziehung

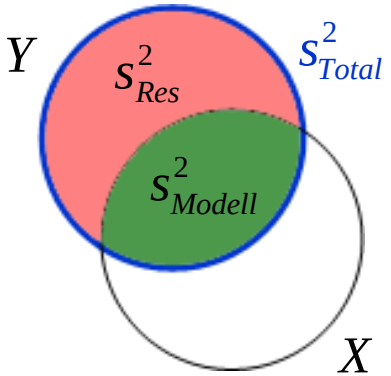
$$\sum_{i=1}^{n}(y_i-\bar{y})=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

auch für die Summe der Abweichungsquadrate (= Quadratsummen = QS) gilt und die Varianzen, die sich ja ergeben durch die Division der Quadratsummen durch $n-1$:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{QS_{Total}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{QS_{Modell}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{QS_{Res}}$$

$$s^2_{Total}=s^2_{Modell}+s^2_{Res}$$

Statt QS findet man in Lehrbüchern auch die Kürzel SS (engl., sum of squares) oder SAQ (=Summe der Abweichungsquadrate).



- Die Gesamtvaribilität (QS_{Total} bzw. s^2_{Total}) zerlegt sich additiv in die durch das lineare Regressionsmodell erklärte Variabilität (QS_{Modell}, s^2_{Modell}) und die nicht erklärte Variabilität (QS_{Res} bzw. s^2_{Res} als Residual- oder Fehlervarianz)

- Die Unterschiede der Personen in der Variable Y lassen sich also zu einem Teil auf Unterschiede in der Variablen X zurückführen, zum Teil aber nicht. Die Fehlervarianz kann mit anderen Einflussgrößen als X zusammenhängen und/oder auch unsystematisch sein.

Produkt-Moment Korrelation

- Daraus ergibt sich, dass ein weiteres geeignetes Maß für die Güte der Vorhersage bzw. die Enge des Zusammenhangs der Anteil der durch das lineare Regressionsmodell erklärten Varianz an der Gesamtvarianz ist (analog auch mit den Varianzen statt den QS darstellbar):

$$R^2 = \frac{QS_{Total} - QS_{Res}}{QS_{Total}} = 1 - \frac{QS_{Res}}{QS_{Total}} = \frac{QS_{Modell}}{QS_{Total}}$$

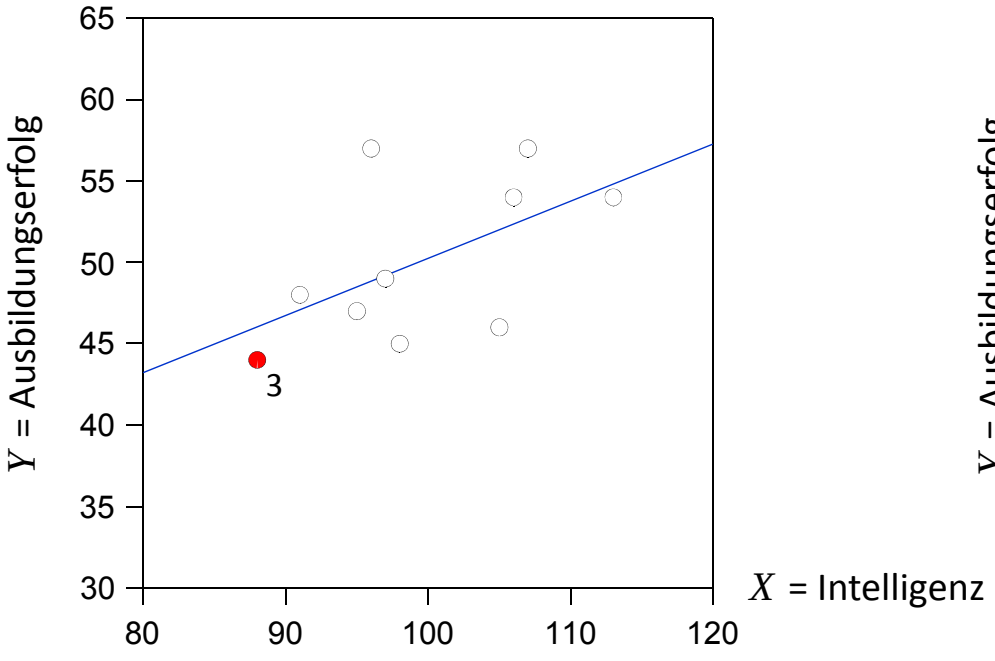
- Dieses Maß wird als auch Determinationskoeffizient R^2 bezeichnet. Er entspricht der quadrierten Produkt-Moment Korrelation (!) und weist folgende Eigenschaften auf:
  - R^2 schwankt zwischen 0 (=kein linearer Zusammenhang) und 1 (=perfekter Zshg.), ist also ein ungerichtetes Maß
  - R^2 ist immer kleinergleich dem Betrag von r (gleich nur bei $r = 0$ und $r = 1$): $R^2 \leq |r|$ z.B. resultiert bei $r = .40$ ein $R^2 = 0.16$; es werden also 16% der Varianz aufgeklärt.
  - R^2 kann interpretiert werden als Anteil der Fehlerreduktion durch den Prädiktor X. QS_{Res} / QS_{Total} gibt an, wie groß der Anteil der nach der Regression verbliebenen Fehlervarianz (= QS_{Res}) an der Varianz bei einer Vorhersage ohne Regression (= QS_{Total}) ist. Entsprechend gibt $R^2 = 1 - QS_{Res} / QS_{Total}$ an, um welchen Anteil die Vorhersagefehler durch die Regression verbessert wurden. Im Beispiel oben wurden die Vorhersagefehler also um 16% reduziert.

Produkt-Moment Korrelation

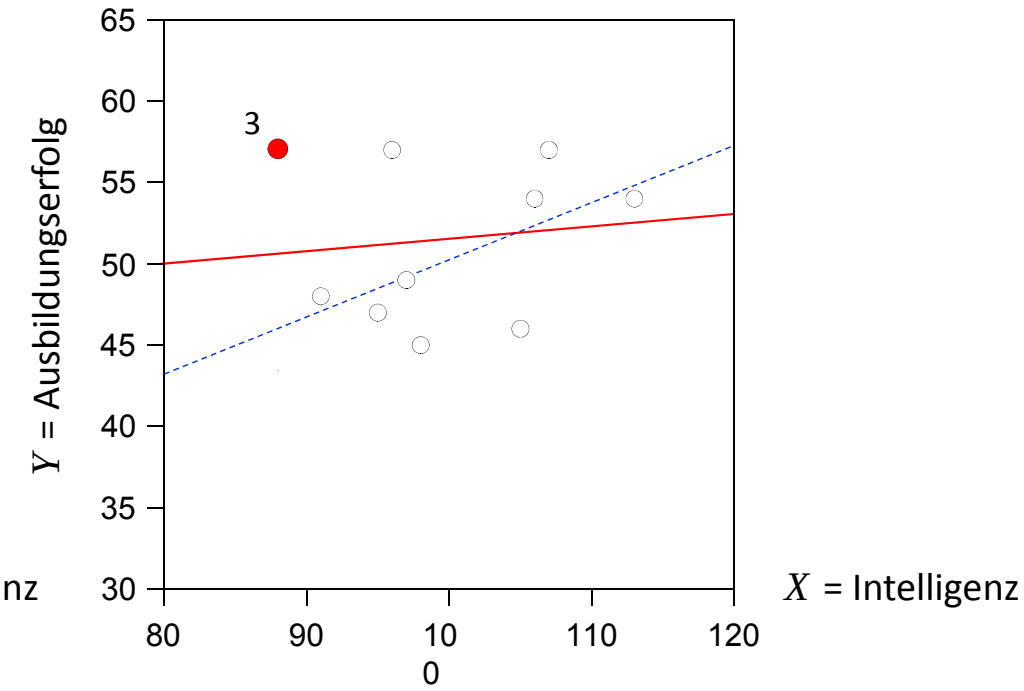
- Die Höhe von Korrelationen kann durch einer Reihe von Faktoren beeinflusst werden, die zu Verzerrungen führen können:
  - Ausreißer
  - Einschränkung in der Varianz von Prädiktor oder Kriterium
  - Nichtlineare Zusammenhänge
  - Heterogene Subgruppen

Produkt-Moment Korrelation

- Korrelationen können (vor allem bei kleinem n) stark durch Ausreißer beeinflusst werden. Je nach Lage der Ausreißer kann dadurch die Korrelation erhöht oder erniedrigt werden.
- Beispiel: Wäre z.B. für Person 3 in unserem Beispiel statt dem korrekten Wert $y_3 = 44$ (wie unten links) der fehlerhafte Wert $y_3 = 57$ eingegeben worden (wie rechts), so vermindert sich die Korrelation von $r = 0.56$ auf $r = 0.13$.



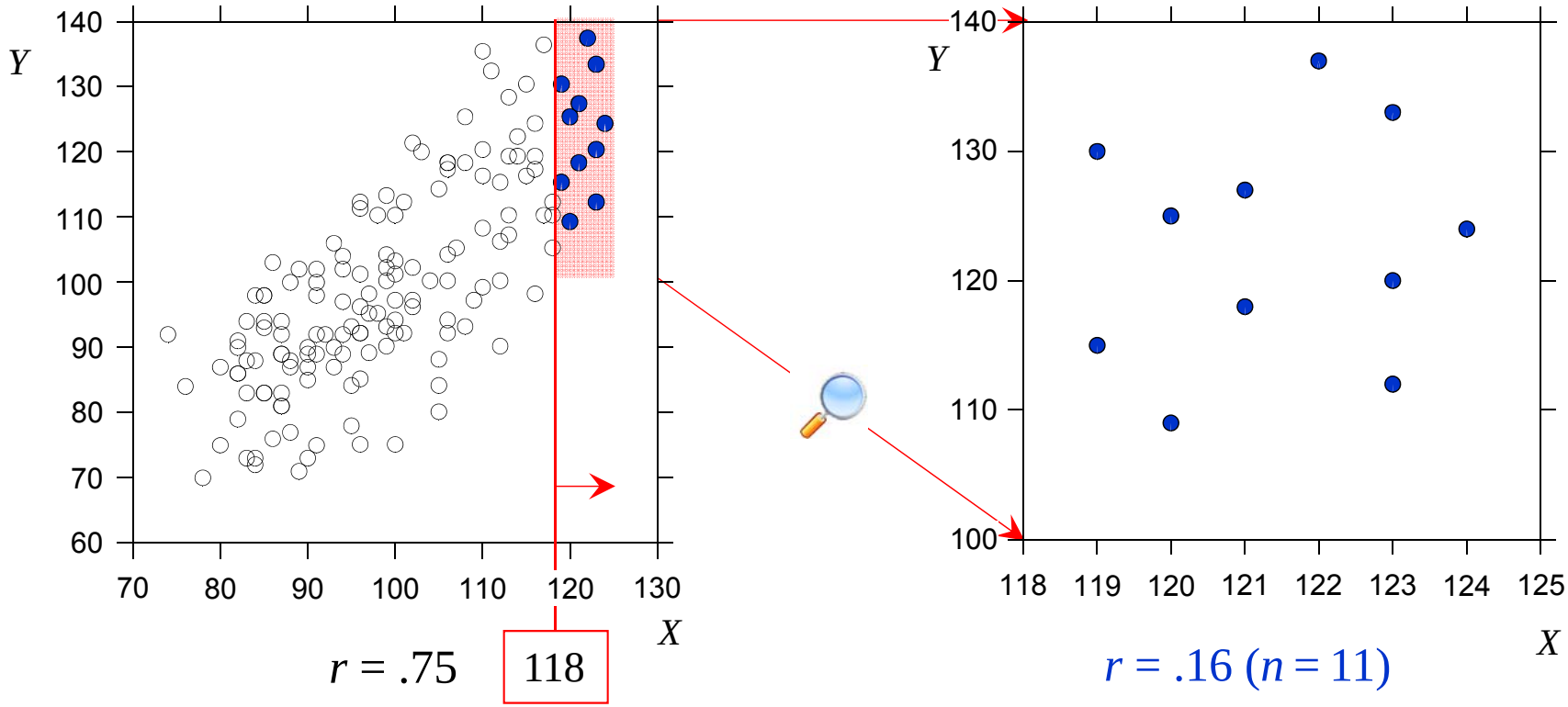
$\hat{y}_i = 15.12 + 0.35 \cdot x_i \qquad r = .56$



$\hat{y}_i = 43.42 + 0.08 \cdot x_i \qquad r = .13$

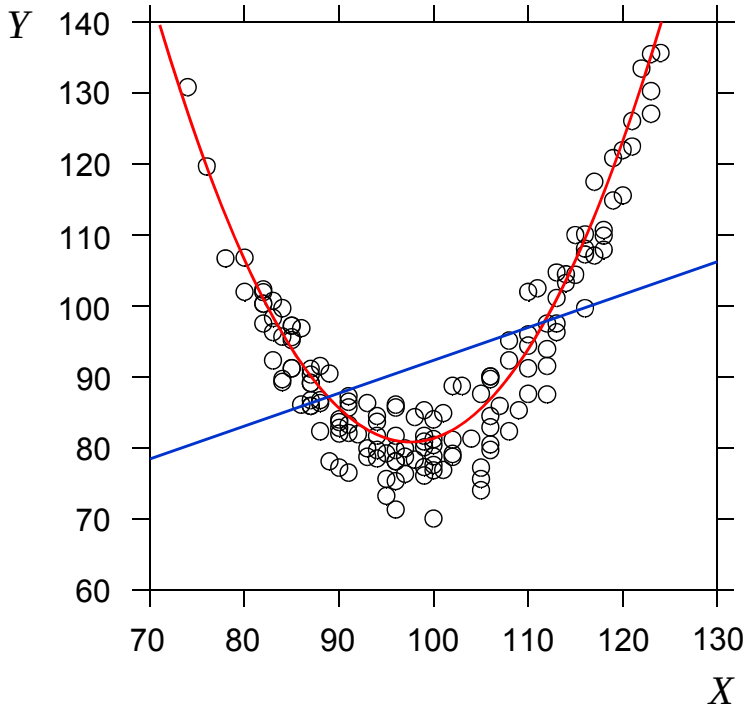
Produkt-Moment Korrelation

- Korrelationen können durch die Einschränkung der Varianz der Variablen (range-restriction) beträchtlich beeinflusst – d.h. meist reduziert – werden.
- Beispiel: Von 150 Bewerbern werden nur die mit einem Intelligenzwert $x > 118$ eingestellt und nach einem Jahr der Berufserfolg (Y) gemessen. Die Korrelation in der untersuchbaren Teilstichprobe rechts unterschätzt den Zusammenhang der Gesamtstichprobe beträchtlich.



Produkt-Moment Korrelation

- Besteht ein nichtlinearer Zusammenhang zwischen X und Y, so ist r das ungeeignete Maß, um diesen Zusammenhang zu quantifizieren.
- Beispiel: Bei untenstehendem U-förmigen Zusammenhang können anhand einer linearen Regression zwar 16% der Varianz erklärt werden. Die Abweichung sind aber systematisch. Durch ein angemessenes nichtlineares Regressionsverfahren lassen sich 92% der Varianz erklären.



$$\hat{y}_i = 45.98 + 0.46 \cdot x_i$$

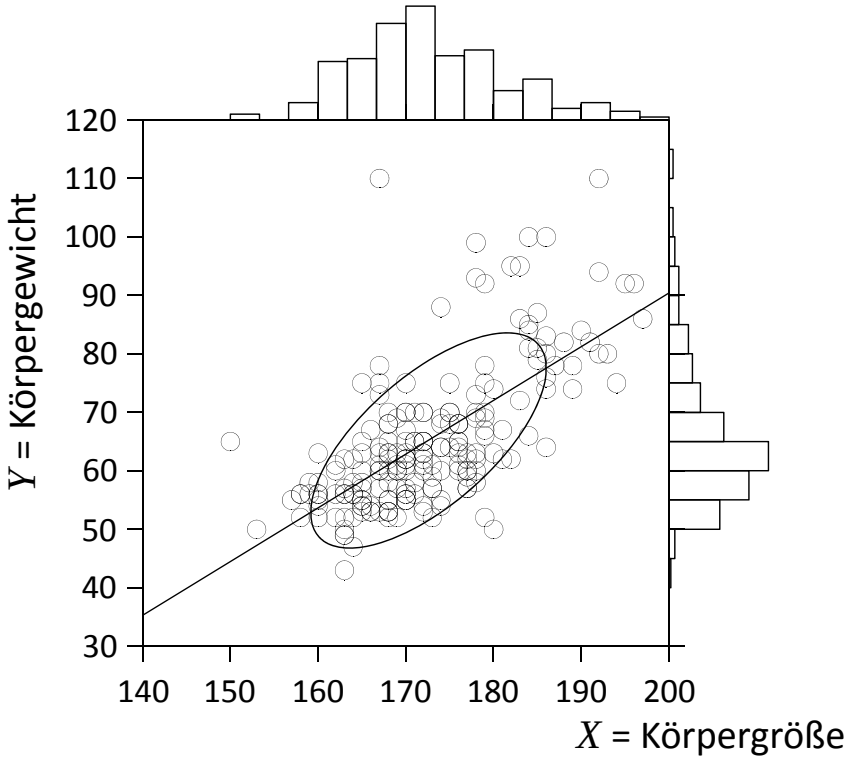
$$r = .40, \quad R^2 = .16$$

$$R^2 = .92$$

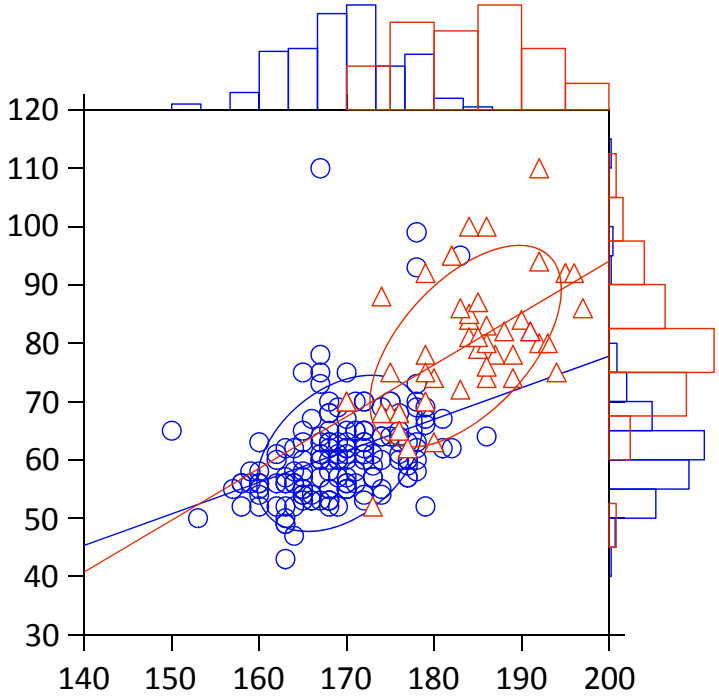
- Preisfrage: Was würde bei einer Varianz-einschränkung durch die Selektion nur hoher oder nur niedriger Werte passieren?

Produkt-Moment Korrelation

- Setzt sich die Stichprobe aus heterogenen Subgruppen zusammen, die sich bezüglich der Enge des Zusammenhangs oder ihrer Lage unterscheiden, kann die Korrelation in der Gesamtstichprobe irreführend sein.
- Beispiel: Zusammenhang zwischen Körpergröße und Körpergewicht in einer Stichprobe von 205 Psychologie-Erstsemestern.



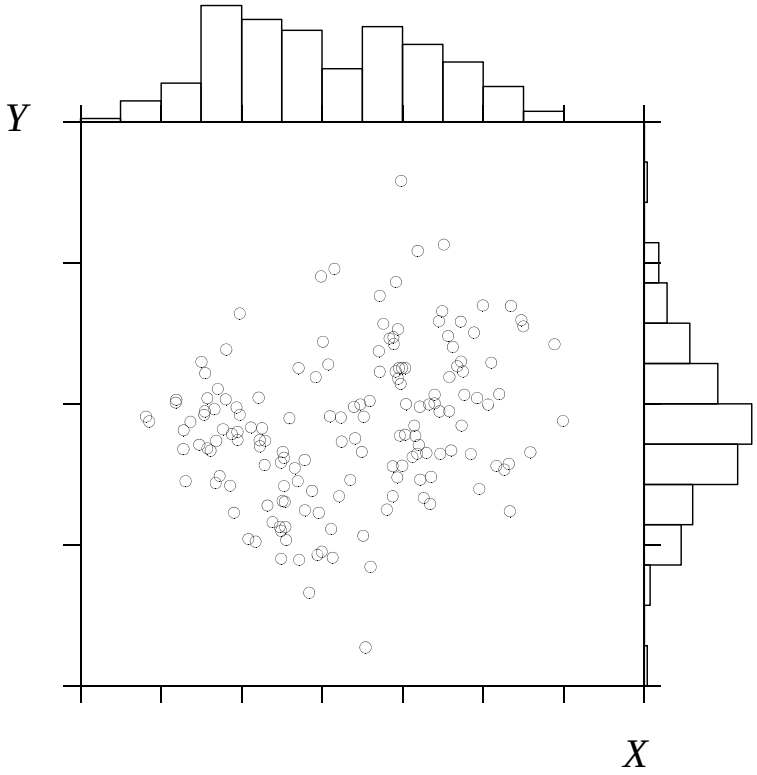
$r = 0.67$ ($n = 205$)



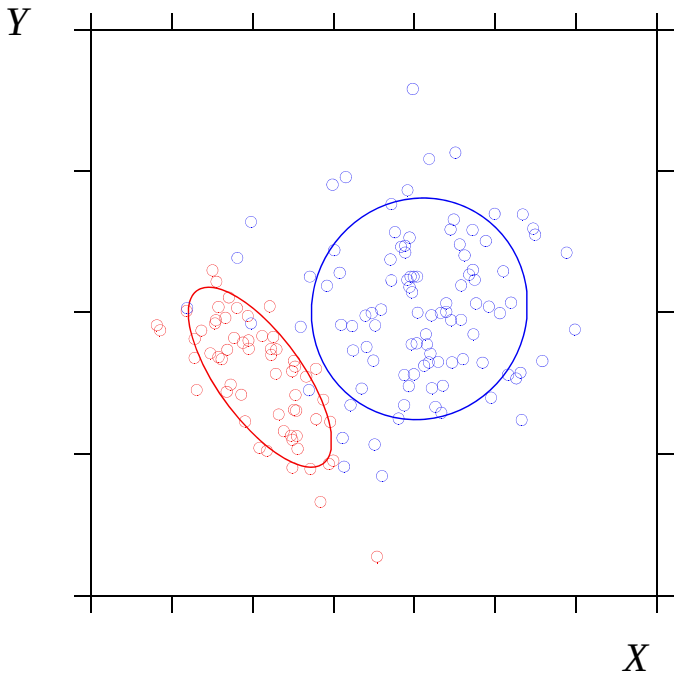
$r = 0.38$ ($n = 157$ ♀) $r = 0.56$ ($n = 46$ ♂)

Produkt-Moment Korrelation

- Im Prinzip sind alle möglichen Effekte durch die Überlagerung von heterogenen Subgruppen möglich. Hier ein weiteres Beispiel:



$r = 0.26$ ($n = 160$)

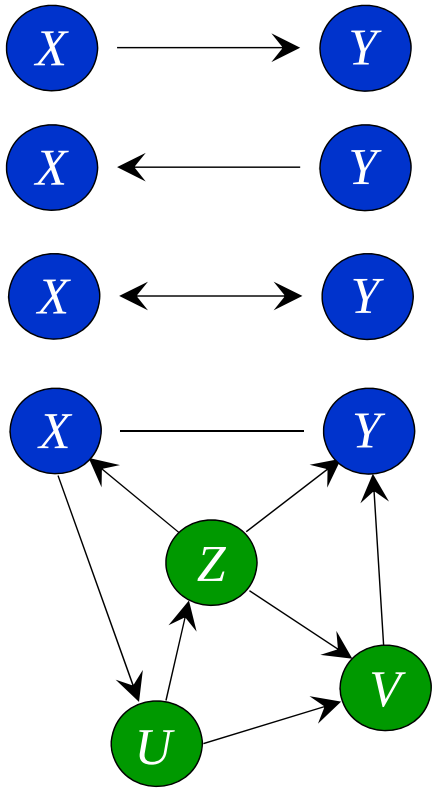


$r = -0.71$ ($n = 60$) $r = 0.04$ ($n = 100$)

Deutung der Korrelation

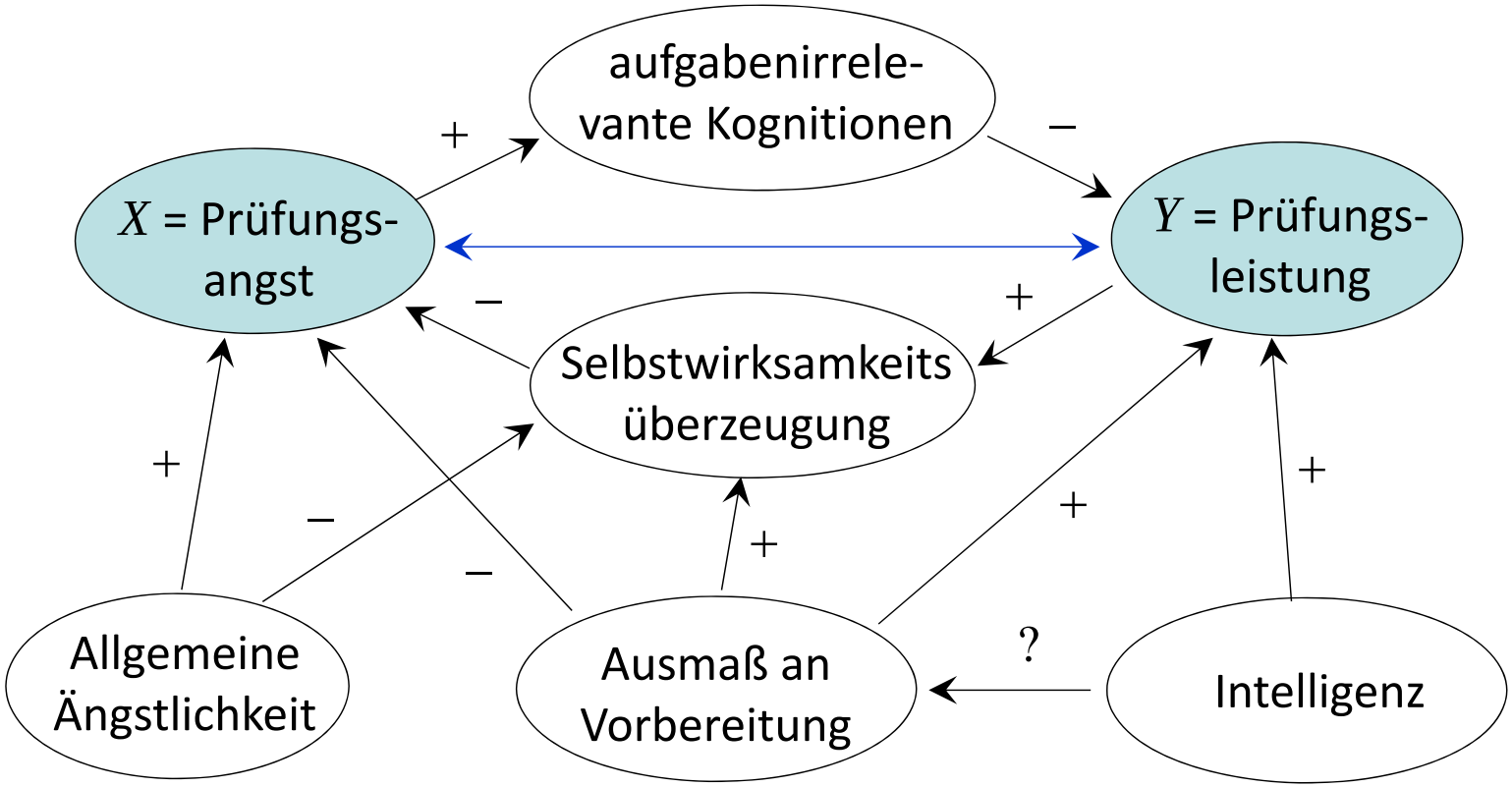
- Einer Korrelation als Maß für die Enge des Zusammenhangs kann man nicht entnehmen, wie die kausale Richtung der Wirkung zwischen X und Y aussieht!
- Beispiel 1: X = Prüfungsangst, Y = Prüfungsleistung, $r = -0.21$.

  Beispiel 2: X = Einkommen, Y = Schuhgröße, $r > 0$.
- Eine Korrelation lässt immer folgende Deutungsmöglichkeiten offen:
  - Variable X beeinflusst Y:
  - Variable Y beeinflusst X:
  - Beide beeinflussen sich gegenseitig:
  - Variablen X und Y werden beide von einer dritten Variablen Z (oder mehreren Variablen) beeinflusst und dadurch entsteht der Zusammenhang:



Deutung der Korrelation

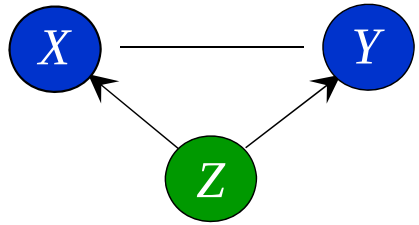
- Meist wird ein Geflecht von Einflüssen eher die Regel sein, am Beispiel:



- Man kann der Korrelation nicht ansehen, welche der Deutungsmöglichkeiten richtig ist. Design zur kausalen Deutung: Experiment.

Partialkorrelation

- Die Partialkorrelation gibt den linearen Zusammenhang zweier Variablen X und Y an, aus dem der lineare Einfluss einer dritten Variablen Z eliminiert wurde.



- Beispiel: Ein positiver Zusammenhang zwischen der Abstraktionsfähigkeit (X) und der sensumotorischen Koordinationsfähigkeit (Y) bei Kindern wird (zumindest zum Teil) gestiftet durch des Einfluss unterschiedlichen Alters (Z).

- Man kann den Einfluss von Z auf statistischem Wege aus dem Zusammenhang von X und Y herausziehen. Dies kann z.B. sinnvoll sein …
  - wenn man vermutet, dass dem Zusammenhang zwischen X und Y eine dritte Variable Z zugrunde liegt (Scheinzusammenhang zwischen X und Y, „spurious correlation“)
  - wenn man zeigen will, dass der Zusammenhang auch bestehen bleibt, wenn man den Einfluss einer dritten Variable kontrolliert hat.

- Eine Alternative zur statistischen Kontrolle ist eine versuchsplanerische Kontrolle (z.B. im obigen Beispiel das Konstanthalten der Störvariable „Alter“).

Partialkorrelation

- Die Partialkorrelation $r_{XY.Z}$ entspricht der Produkt-Moment Korrelation zwischen den Residuen, die entstehen, wenn X und Y jeweils per linearer Regression aufgrund von Z vorhergesagt werden:

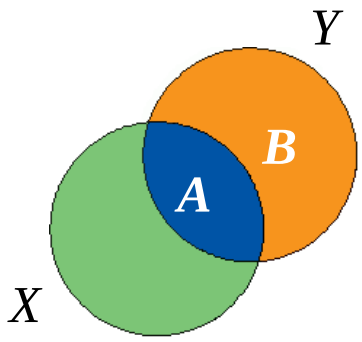
$$r_{XY.Z} = r\left(x_i - \hat{x}_i, y_i - \hat{y}_i\right)$$

$\hat{x}_i = a + b \cdot z_i$ $\quad$ $\hat{y}_i = c + d \cdot z_i$

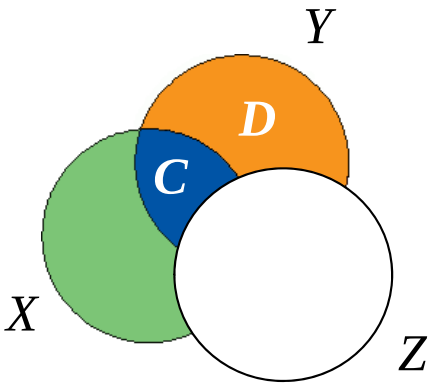
- Statt zwei Regressionen und eine Korrelation zu berechnen, kann man einfacher direkt in folgende Formel einsetzen (Dazu schreiben wir immer die beiden Variablen als Subskripte von r; r_{XY} bezeichnet also die Korrelation zwischen X und Y.):

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1 - r_{XZ}^2} \cdot \sqrt{1 - r_{YZ}^2}}$$

- In Venn-Diagramm-Darstellung der Varianzen der drei Variablen X, Y und Z lässt sich der Zusammenhang wie folgt darstellen:



$r_{XY}^2 = A/(A+B)$



$r_{XY.Z}^2 = C/(C+D)$

Partialkorrelation

- Beispiel (fiktiv aus Bortz & Schuster, 2010): Bei 12 Kindern wurde der Zusammenhang der Abstraktionsfähigkeit (= X) und der sensumotorischen Koordination (= Y) untersucht. Es liegt nahe, dass der positive Zusammenhang von $r_{XY} = .89$, zumindest zum Teil, durch das Alter (= Z) zustande kommt.

  Die Korrelationen beider Variablen mit dem Alter betragen $r_{XZ} = .77$, $r_{YZ} = .80$.

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{1 - r_{XZ}^2} \cdot \sqrt{1 - r_{YZ}^2}}$$

$$= \frac{0.89 - 0.77 \cdot 0.80}{\sqrt{1 - 0.77^2}\sqrt{1 - 0.80^2}} = .72$$

- Der Zusammenhang verringert sich durch das Herauspartialisieren des Alters von $r_{XY} = .89$ auf $r_{XY.Z} = .72$.

Vp-Nr.	x_i	y_i	z_i	$x_i - \hat{x}_i$	$y_i - \hat{y}_i$
1	9	8	6	1.058	0.609
2	11	12	8	0.565	1.768
3	13	14	9	1.319	2.348
4	13	13	9	1.319	1.348
5	14	14	10	1.072	0.928
6	9	8	7	-0.188	-0.812
7	10	9	8	-0.435	-1.232
8	11	12	9	-0.681	0.348
9	10	8	8	-0.435	-2.232
10	8	9	7	-1.188	0.188
11	13	14	10	0.072	0.928
12	7	7	6	-0.942	-0.391
13	9	10	10	-3.928	-3.072
14	13	12	10	0.072	-1.072
15	14	12	9	2.319	0.348

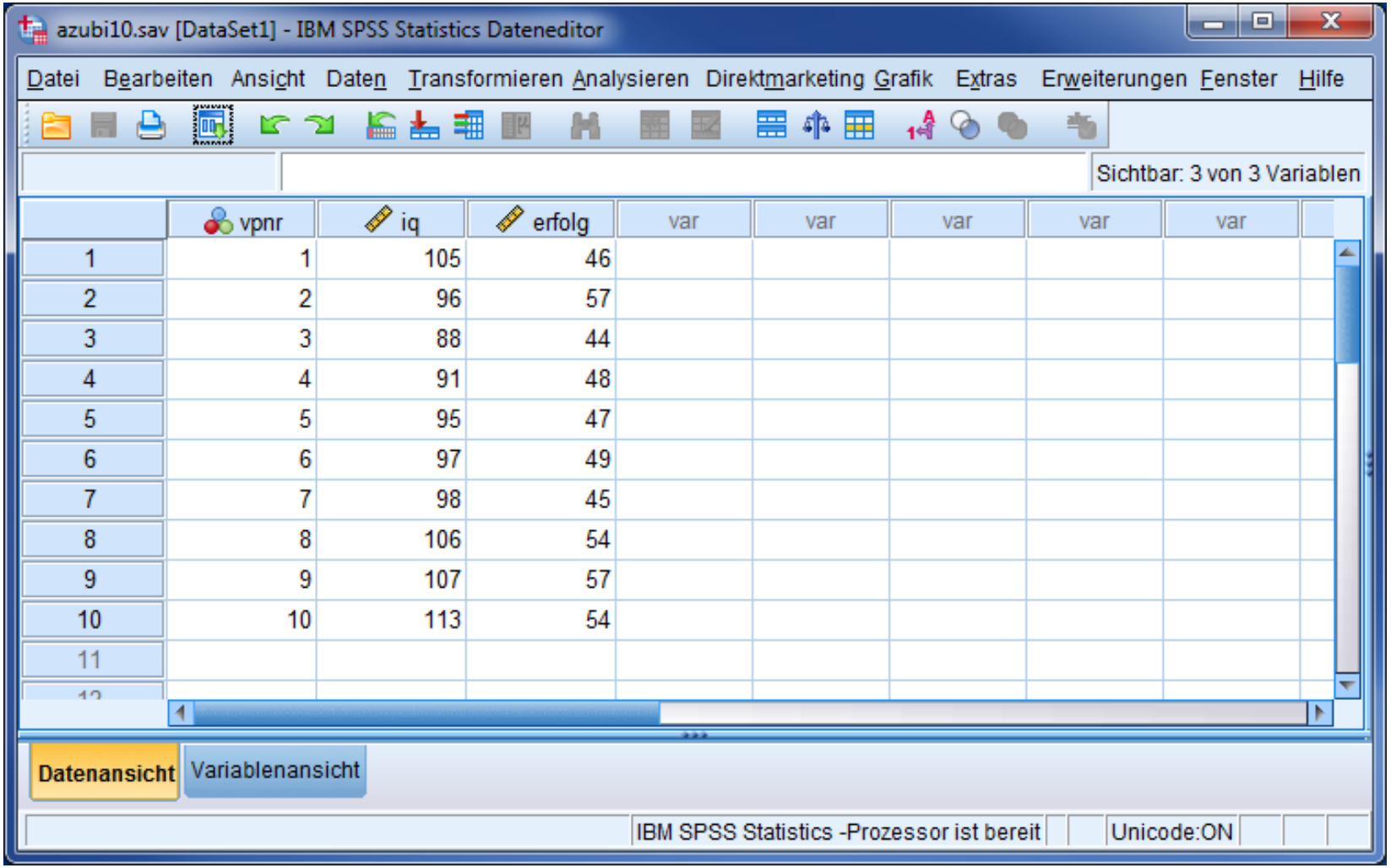
$$x_i = 0.464 + 1.246 \cdot z_i + e_i$$

$$y_i = -1.130 + 1.420 \cdot z_i + e_i$$

$r = .72$

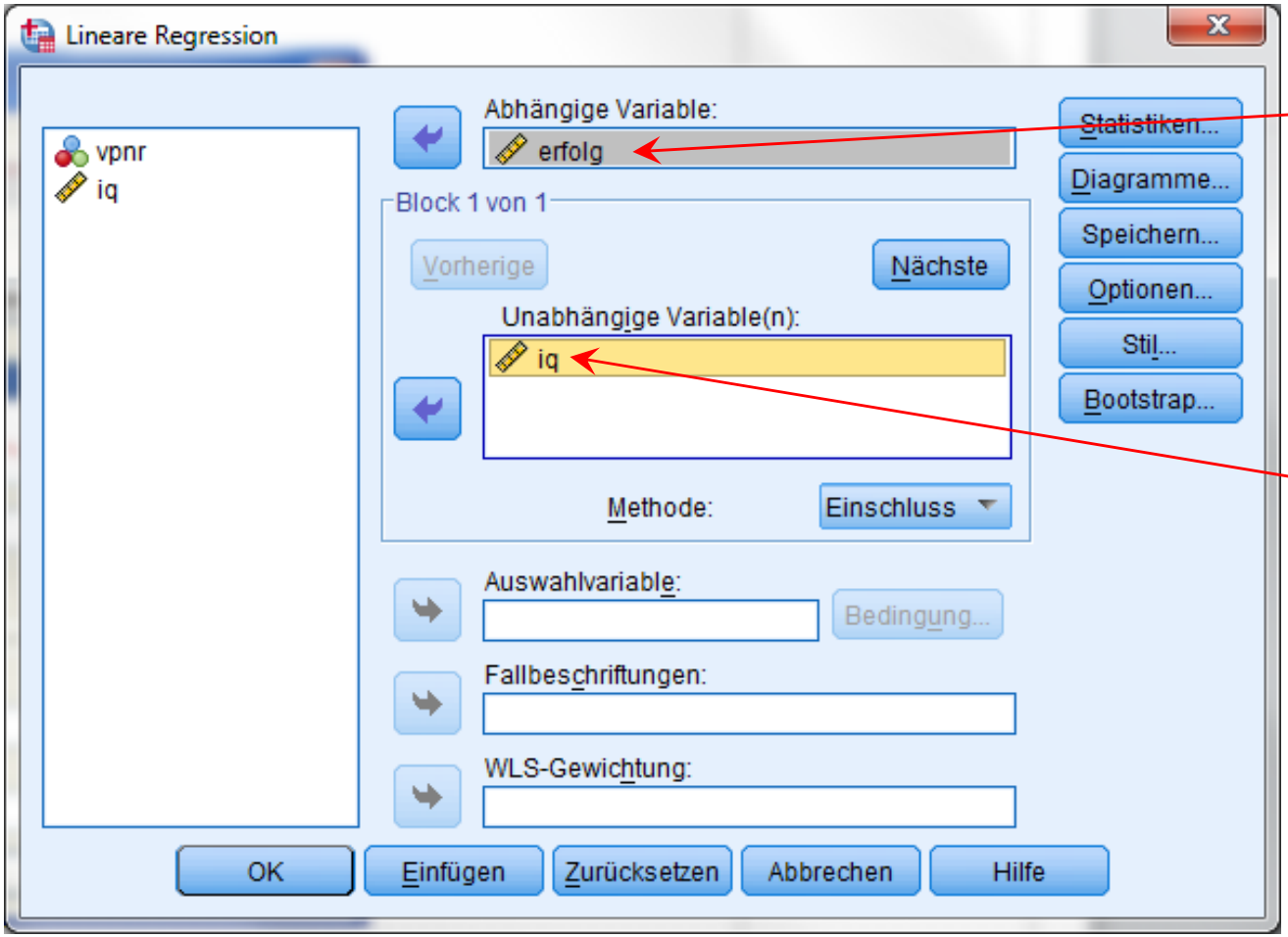
SPSS: Bivariate lineare Regression & Korrelation

- Der Beispieldatensatz enthält die Variablen Versuchspersonen-Nummer (VPNR), Ergebnisse im Intelligenztest (IQ) und Ausbildungserfolg (ERFOLG) für $n = 10$ Azubis:



Regression in SPSS

- Die lineare Regressionsgleichung erhält man unter `Analysieren/Regression/Linear`…. (Prozedur: *Regression*)



Die Kriteriumsvariable Y ist unter „Abhängige Variable“ einzugeben.

Die Prädiktorvariable X ist unter „Unabhängige Variable“ einzugeben.

Regression

... (Tabellen weggelassen) ...

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,557[a]	,311	,224	4,363

a. Einflußvariablen : (Konstante), iq

In der Spalte „Standardfehler des Schätzers“ der Tabelle „Modellzusammenfassung“ findet sich der Standardschätzfehler $s_E = 4.36$

Koeffizienten[a]

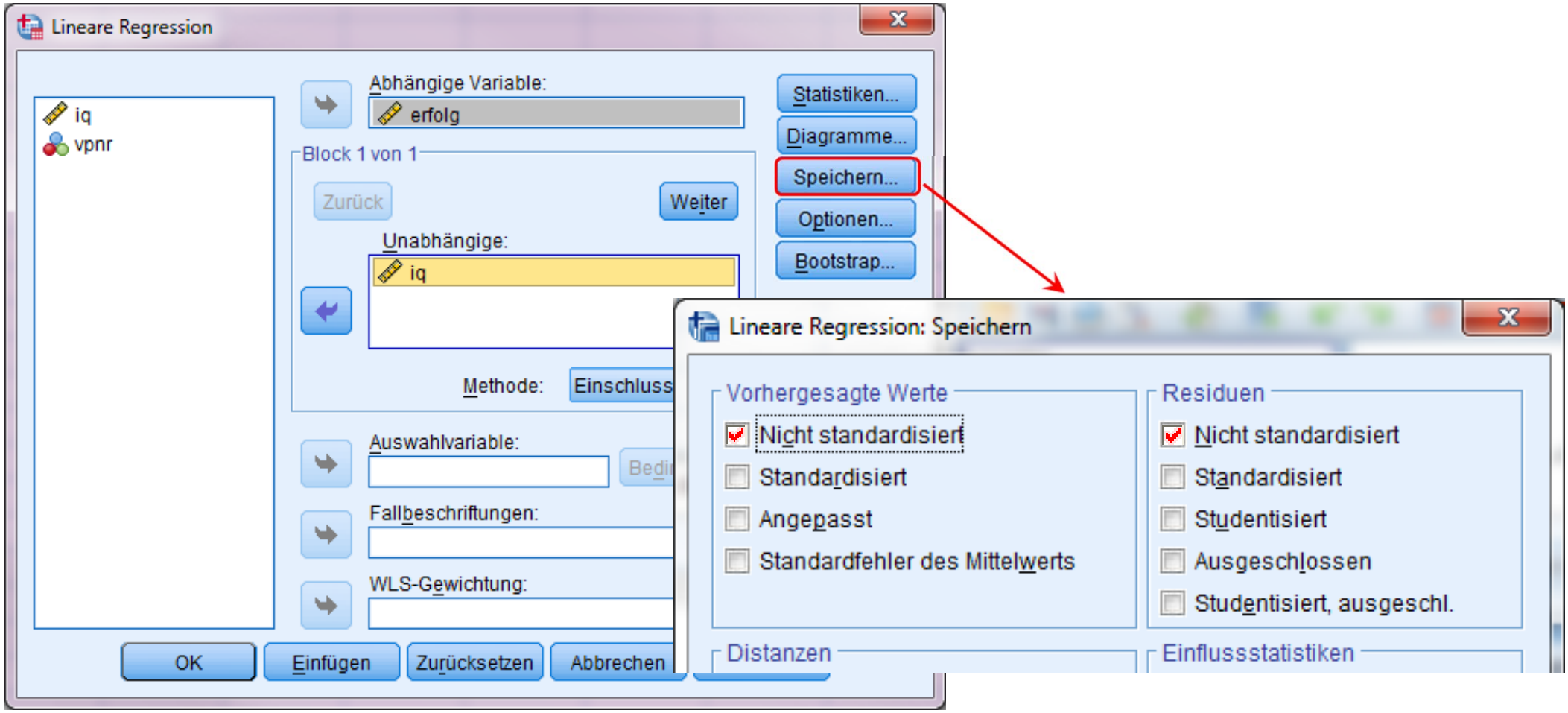
Modell		Nicht standardisierte Koeffizienten: Regressionskoeffizient B	Nicht standardisierte Koeffizienten: Standardfehler	Standardisierte Koeffizienten: Beta	T	Sig.
1	(Konstante)	15,122	18,474		,819	,437
	iq	,351	,185	,557	1,899	,094

a. Abhängige Variable: erfolg

In der Spalte „RegressionskoeffizientB“ der Tabelle „Koeffizienten“ finden sich die beiden Regressionsgewichte a in der Zeile „(Konstante)“ und b in der Zeile mit dem Namen der Prädiktorvariablen (hier: `IQ`).

$$\hat{y}_i = a + b \cdot x_i = 15.122 + 0.351 \cdot x_i$$

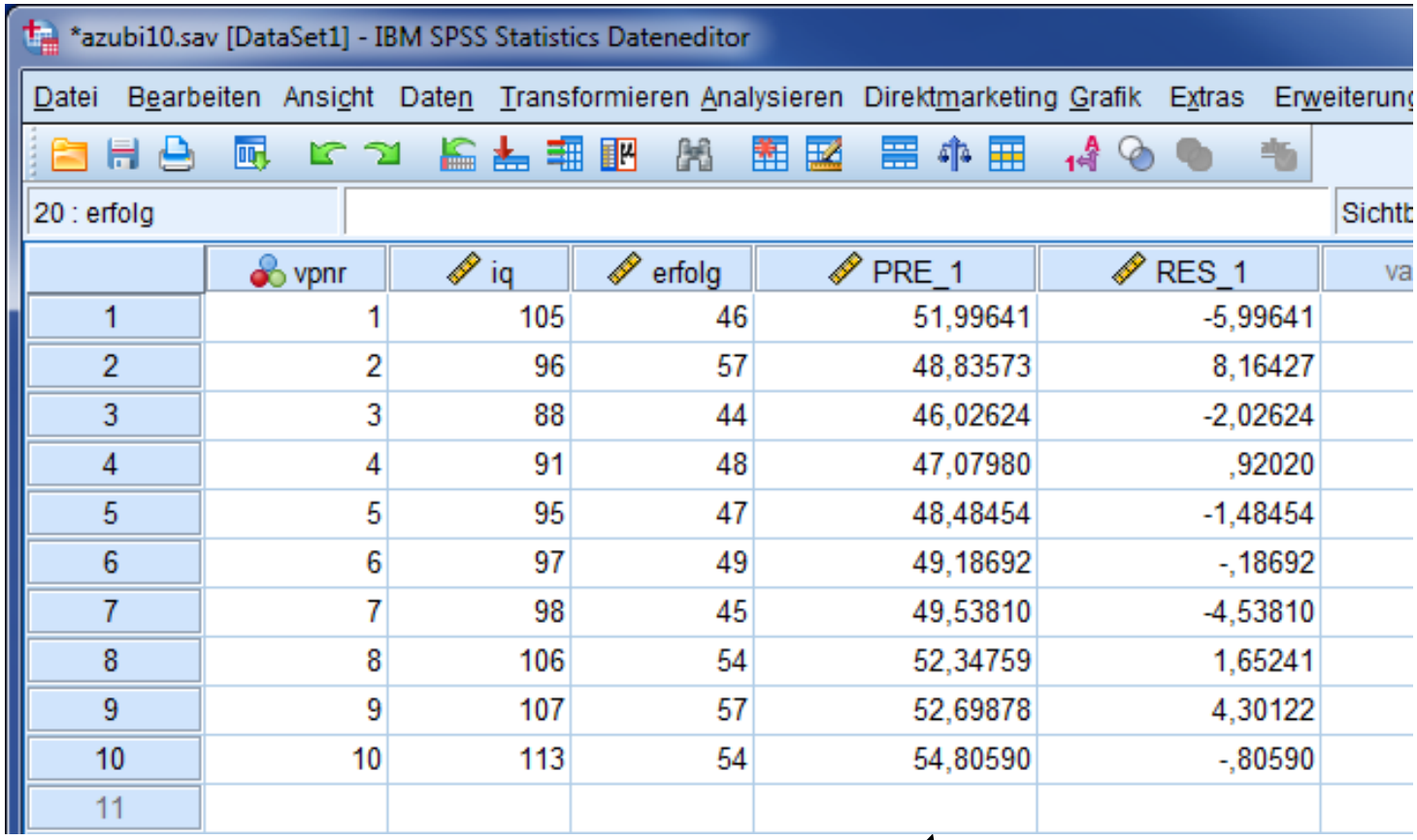
Regression in SPSS



Nach Klicken von (Speichern) können zusätzlich die vorhergesagten Werte $\hat{y}_i$ („Vorhergesagte Werte: Nicht standardisiert“) sowie die Residuen $e_i = y_i - \hat{y}_i$ („Residuen: Nicht standardisiert“) als zusätzliche, neue Variablen an die Datendatei angefügt werden.

Regression in SPSS

- Die Spalte `PRE_1` enthält dann die vorhergesagten Werte, die Spalte `RES_1` die Residuen.



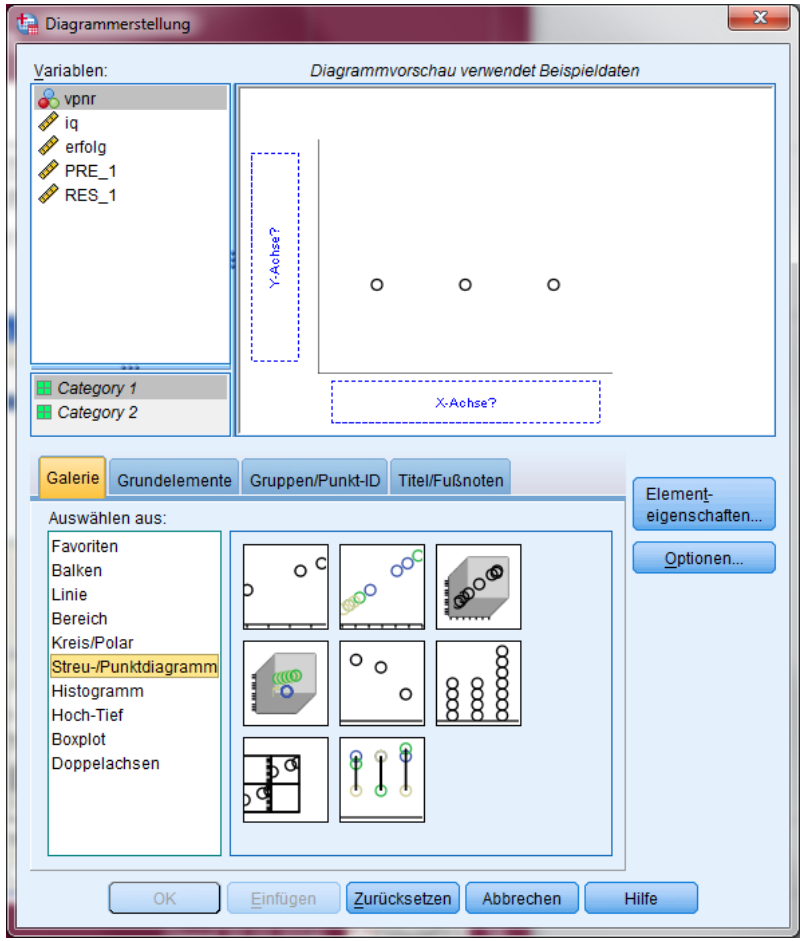
	vpnr	iq	erfolg	PRE_1	RES_1	va
1	1	105	46	51,99641	-5,99641	
2	2	96	57	48,83573	8,16427	
3	3	88	44	46,02624	-2,02624	
4	4	91	48	47,07980	,92020	
5	5	95	47	48,48454	-1,48454	
6	6	97	49	49,18692	-,18692	
7	7	98	45	49,53810	-4,53810	
8	8	106	54	52,34759	1,65241	
9	9	107	57	52,69878	4,30122	
10	10	113	54	54,80590	-,80590	
11						

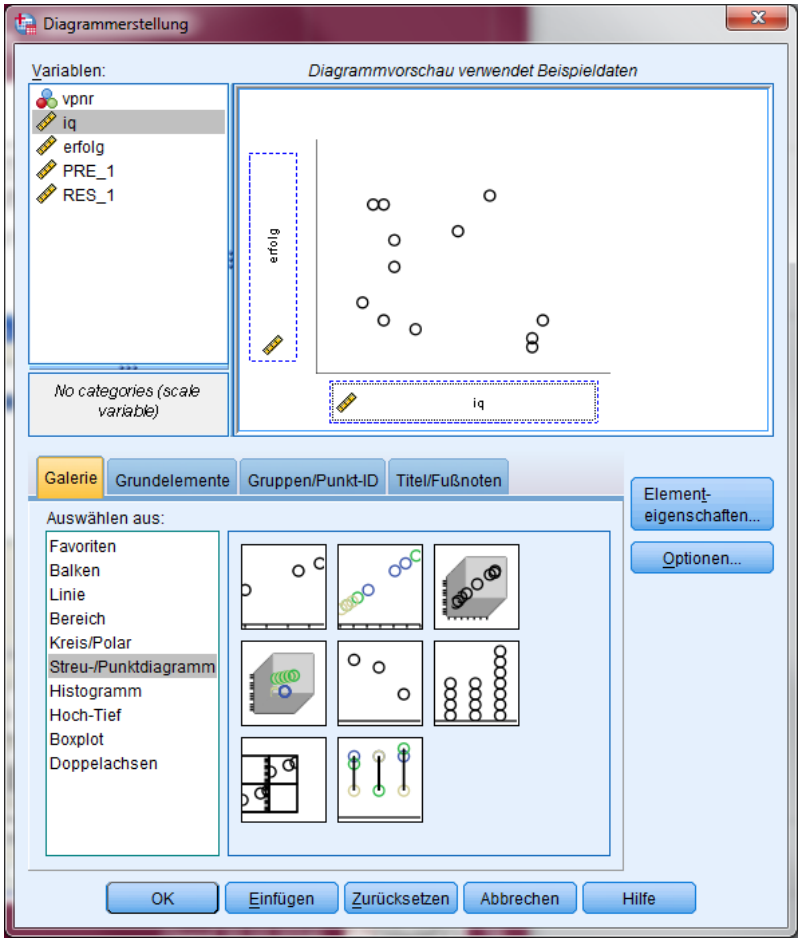
$$\hat{y}_i = a + b \cdot x_i = 15.122 + 0.351 \cdot x_i$$

$$e_i = y_i - \hat{y}_i$$

SPSS: Streudiagramm

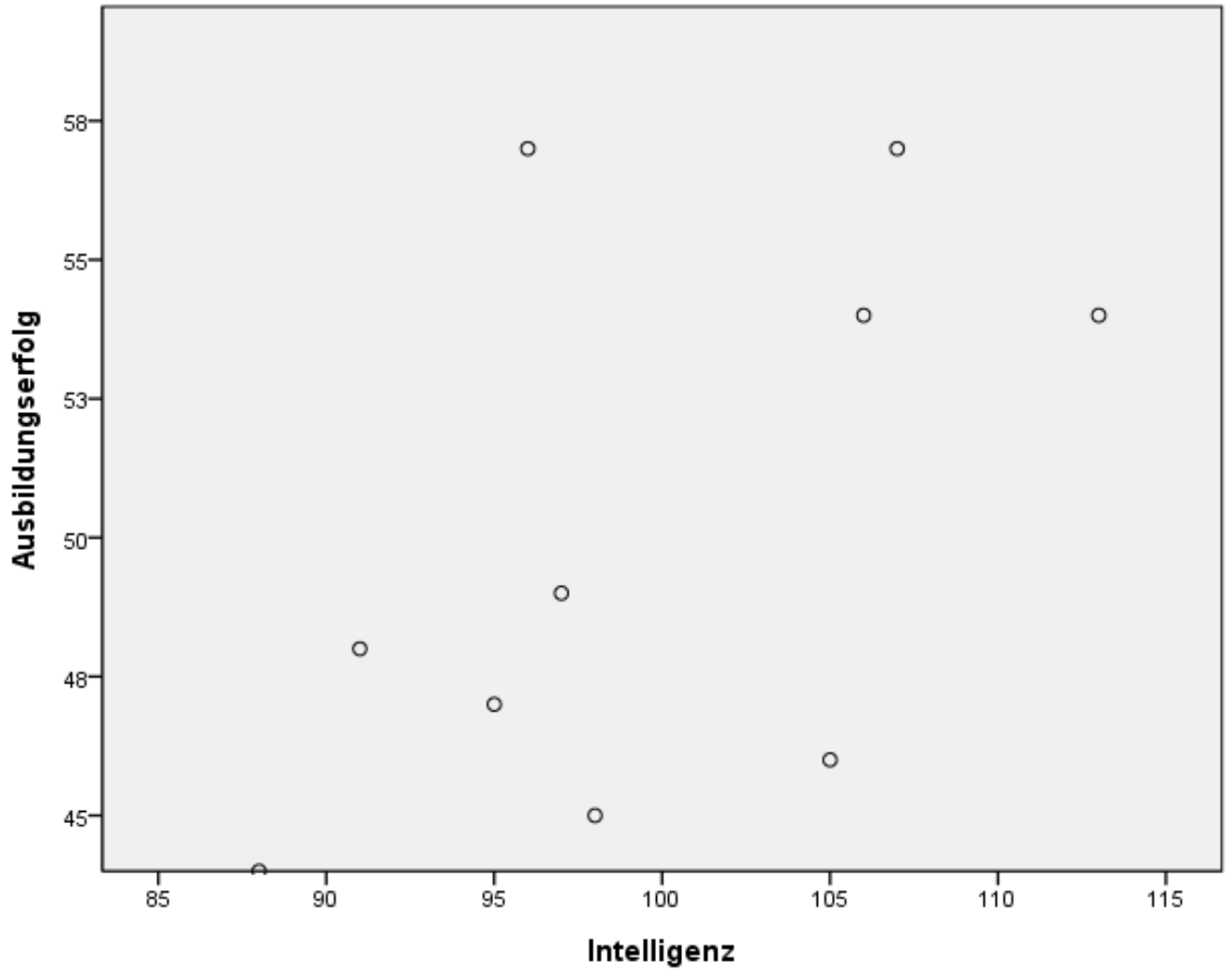
- Ein Streudiagramm erhält man unter `Grafik/Diagrammerstellung…`, wenn man dort im Reiter Galerie die Option „Streu-/Punktdiagramm“ anwählt und das einfache Diagramm  in die Diagrammvorschau zieht. Anschließend sind die Variablen `IQ` auf X-Achse? und ERFOLG auf Y-Achse? zu ziehen.





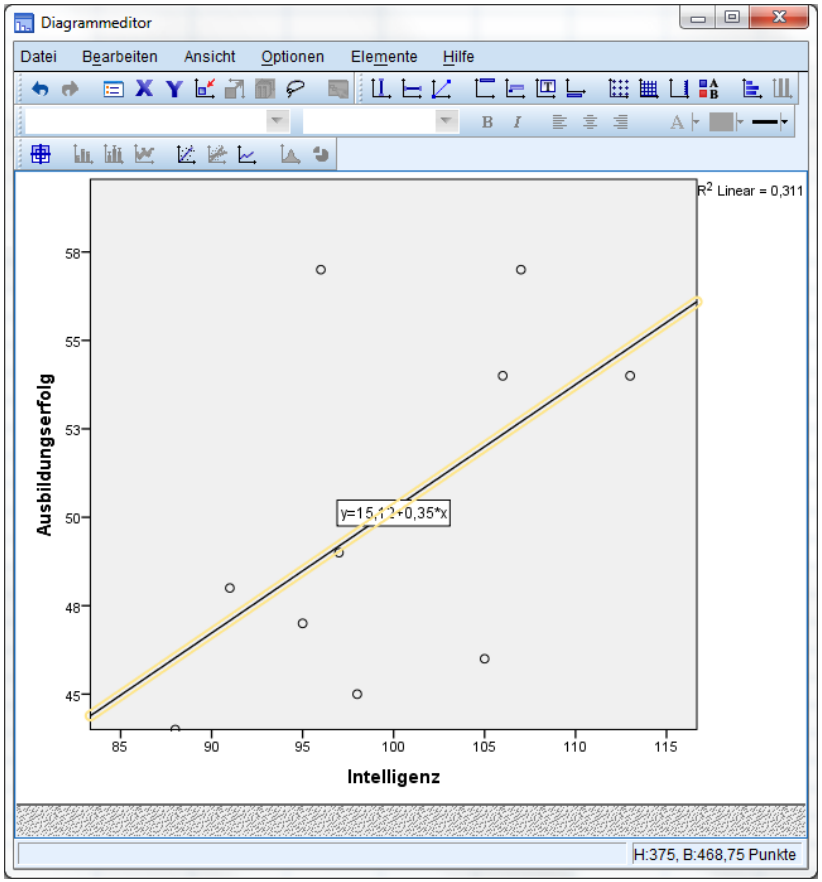
SPSS: Streudiagramm

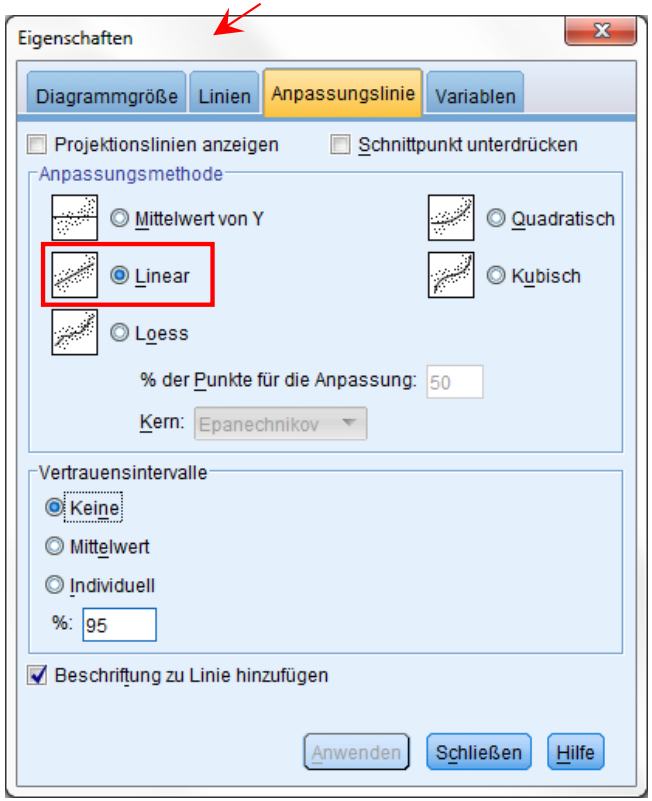
- Es resultiert standardmäßig das folgende Streudiagramm:



SPSS: Streudiagramm

- Nach dem Wechsel in den Diagramm-Editor kann die lineare Regressionsgerade (und die Gleichung) hinzugefügt werden, indem man die Option `Elemente/Anpassungslinie_bei Gesamtsumme` oder das Icon  anklickt. In dem Dialog [Eigenschaften] können verschiedene Anpassungsmethoden eingestellt werden; „linear“ ist voreingestellt.

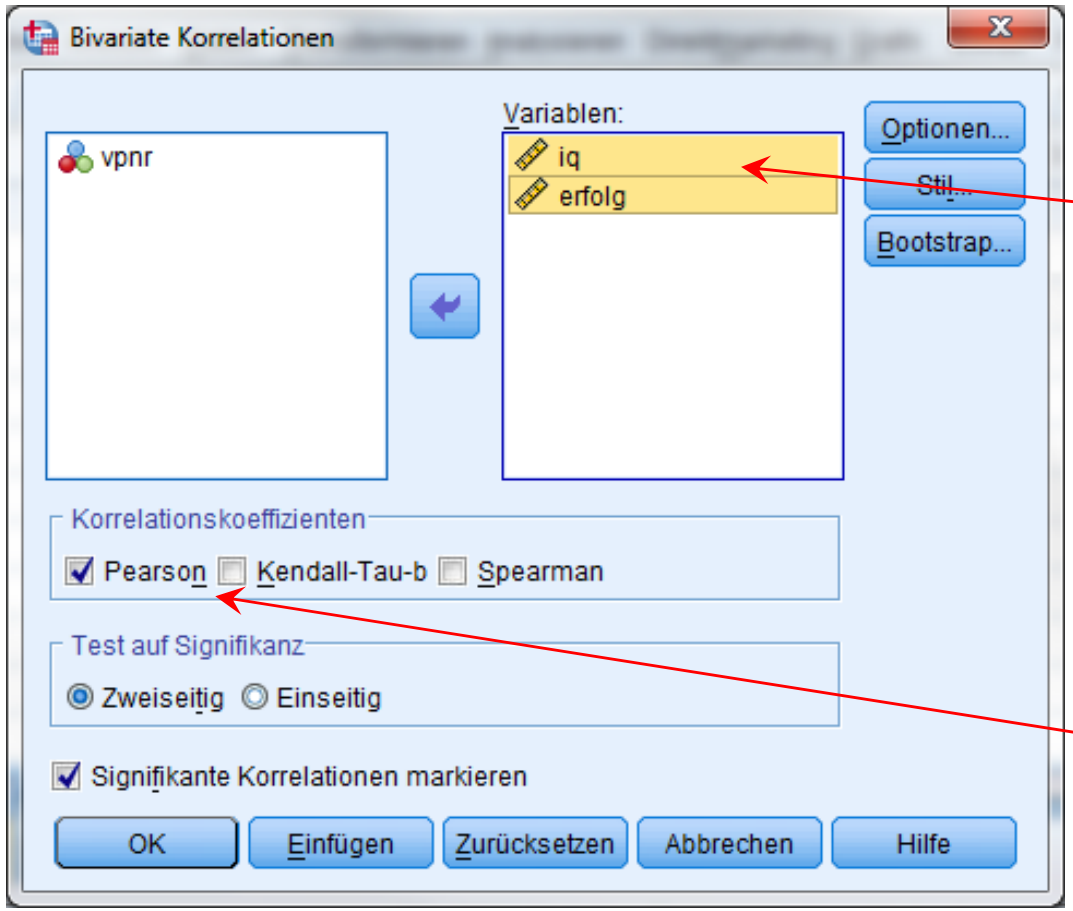




Hinweis: Nach Doppelklicken auf die Gleichung kann man in der Eigenschaft „Bezugslinie“ die Gleichung ausblenden, indem man die Option „Beschriftung zu Linie hinzufügen“ deaktiviert.

Produkt-Moment Korrelation in SPSS

- Die Produkt-Moment Korrelation *r* erhält man unter `Analysieren/Korrelation/Bivariat…` (Prozedur *CORRELATIONS*):

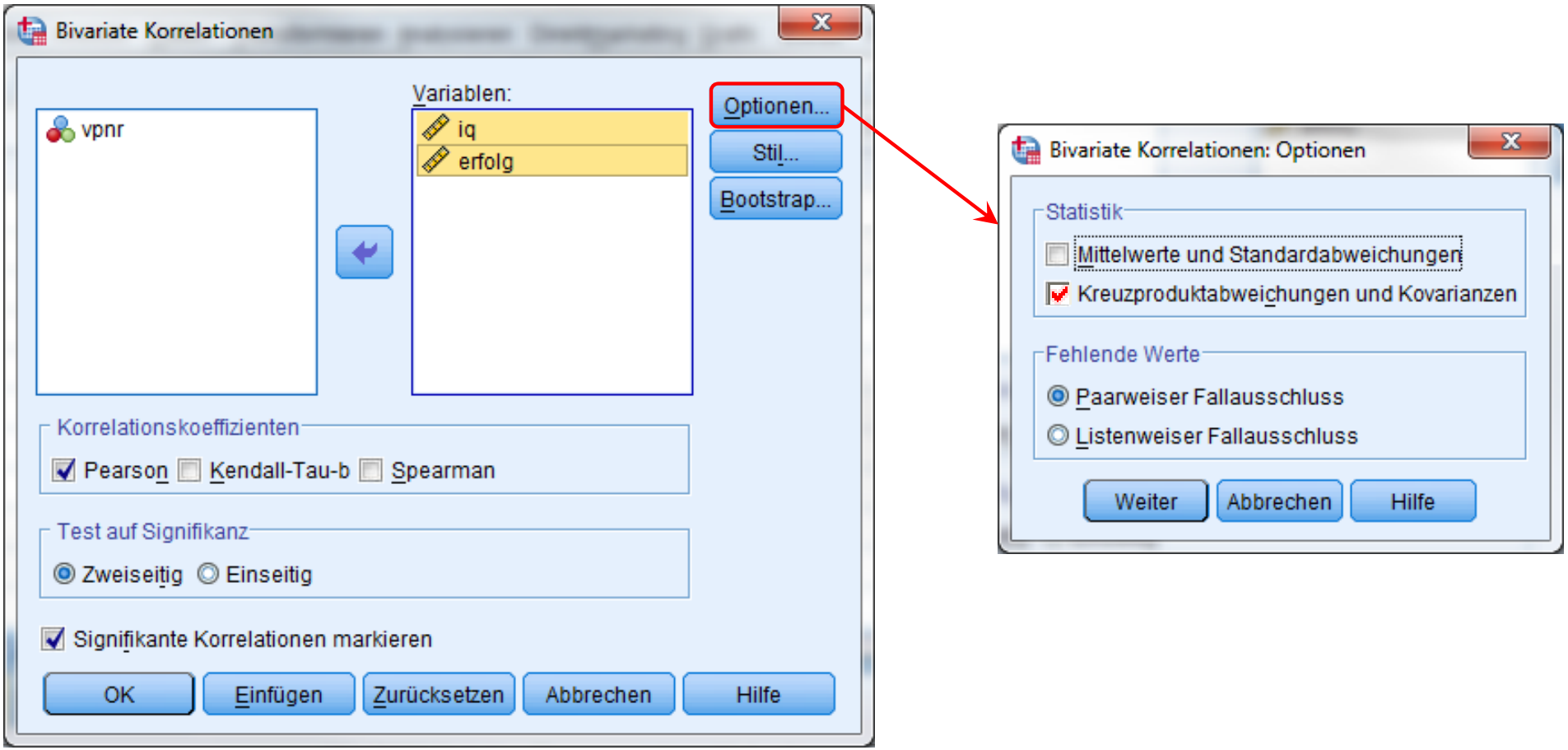


Unter „Variablen“ sind die zu korrelierenden Variablen anzugeben. Werden mehr als zwei Variablen angewählt, so werden die Korrelationen für alle Variablenpaare bestimmt.

Die Produkt-Moment Korrelation nach Pearson ist schon voreingestellt.

Produkt-Moment Korrelation in SPSS

- Soll (zusätzlich) die Kovarianz ausgegeben werden, so ist nach Klicken von (`Optionen`) unter »Statistiken« die Option „Kreuzproduktabweichungen und Kovarianzen“ zu aktivieren.



Produkt-Moment Korrelation in SPSS

Korrelation

Y

Korrelationen

		iq	erfolg
iq	Korrelation nach Pearson	1	,557
	Signifikanz (2-seitig)		,094
X →	Quadratsummen und Kreuzprodukte	556,400	195,400
	Kovarianz	61,822	21,711
	N	10	10
erfolg	Korrelation nach Pearson	,557	1
	Signifikanz (2-seitig)	,094	
	Quadratsummen und Kreuzprodukte	195,400	220,900
	Kovarianz	21,711	24,544
	N	10	10

$r(X,Y) = 0.557$

Kreuzprodukt = 195.4

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} = \frac{195.4}{9} = 21.711$$

Wie man sieht, gilt

- $r(X, Y) = r(Y, X)$
- $cov(X, Y) = cov(Y, X)$

Produkt-Moment Korrelation in SPSS

Korrelation

X

Korrelationen

		iq	erfolg
iq	Korrelation nach Pearson	1	,557
	Signifikanz (2-seitig)		,094
X	Quadratsummen und Kreuzprodukte	556,400	195,400
	Kovarianz	61,822	21,711
	N	10	10
erfolg	Korrelation nach Pearson	,557	1
	Signifikanz (2-seitig)	,094	
	Quadratsummen und Kreuzprodukte	195,400	220,900
	Kovarianz	21,711	24,544
	N	10	10

$cov(X, X) = s^2(X)=61.822$

Wie man sieht, gilt ferner

- $r(X, X) = r(Y, Y) = 1$
- $cov(X, X) = s^2(X)$; $cov(Y, Y) = s^2(Y)$

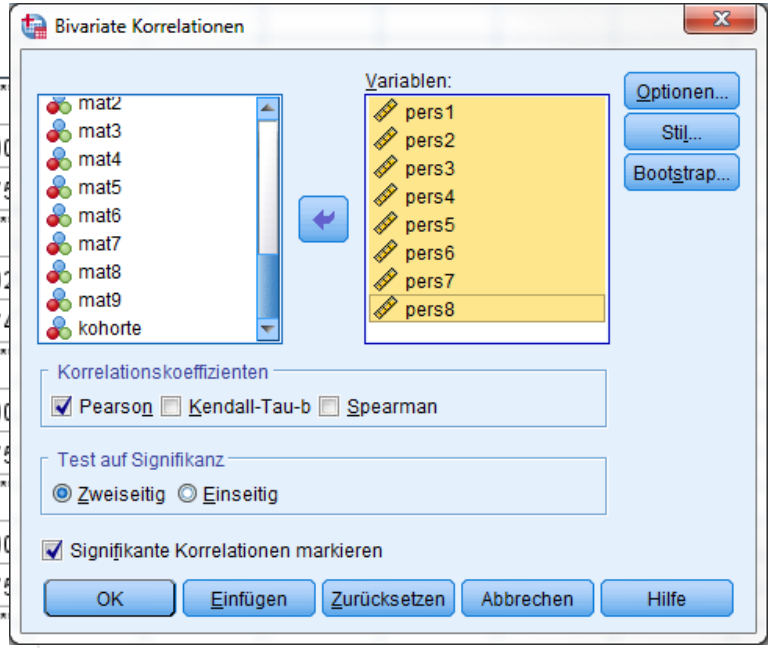
$$cov(X, X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (x_i - \bar{x})}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = s^2(X)$$

Produkt-Moment Korrelation in SPSS

Korrelationen

		pers1	pers2	pers3	pers4	pers5	pers6	pers7	
pers1	Korrelation nach Pearson	1	-,146**	-,030	-,212**	,226**	,213**	-,192**	
	Signifikanz (2-seitig)		,001	,512	,000	,000	,000	,000	
	N	476	475	476	476	474	476	475	
pers2	Korrelation nach Pearson	-,146**	1	,366**	,323**	-,215**	-,275**	,140**	
	Signifikanz (2-seitig)	,001		,000	,000	,000	,000	,002	
	N	475	475	475	475	474	475	474	
pers3	Korrelation nach Pearson	-,030	,366**	1	,310**	-,421**	-,330**	,204**	
	Signifikanz (2-seitig)	,512	,000		,000	,000	,000	,000	
	N	476	475	476	476	474	476	475	
pers4	Korrelation nach Pearson	-,212**	,323**	,310**	1	-,417**	-,438**	,218**	
	Signifikanz (2-seitig)	,000	,000	,000		,000	,000	,000	
	N	476	475	476	476	474	476	475	
pers5	Korrelation nach Pearson	,226**	-,215**	-,421**	-,417**	1	,565**	-,191**	
	Signifikanz (2-seitig)	,000	,000	,000	,000		,000	,000	,000
	N	474	474	474	474	474	474	473	468
pers6	Korrelation nach Pearson	,213**	-,275**	-,330**	-,438**	,565**	1	-,237**	,435**
	Signifikanz (2-seitig)	,000	,000	,000	,000	,000		,000	,000
	N	476	475	476	476	474	476	475	470
pers7	Korrelation nach Pearson	-,192**	,140**	,204**	,218**	-,191**	-,237**	1	-,175**
	Signifikanz (2-seitig)	,000	,002	,000	,000	,000	,000		,000
	N	475	474	475	475	473	475	475	470
pers8	Korrelation nach Pearson	,261**	-,221**	-,174**	-,374**	,357**	,435**	-,175**	1
	Signifikanz (2-seitig)	,000	,000	,000	,000	,000	,000	,000	
	N	470	469	470	470	468	470	470	470

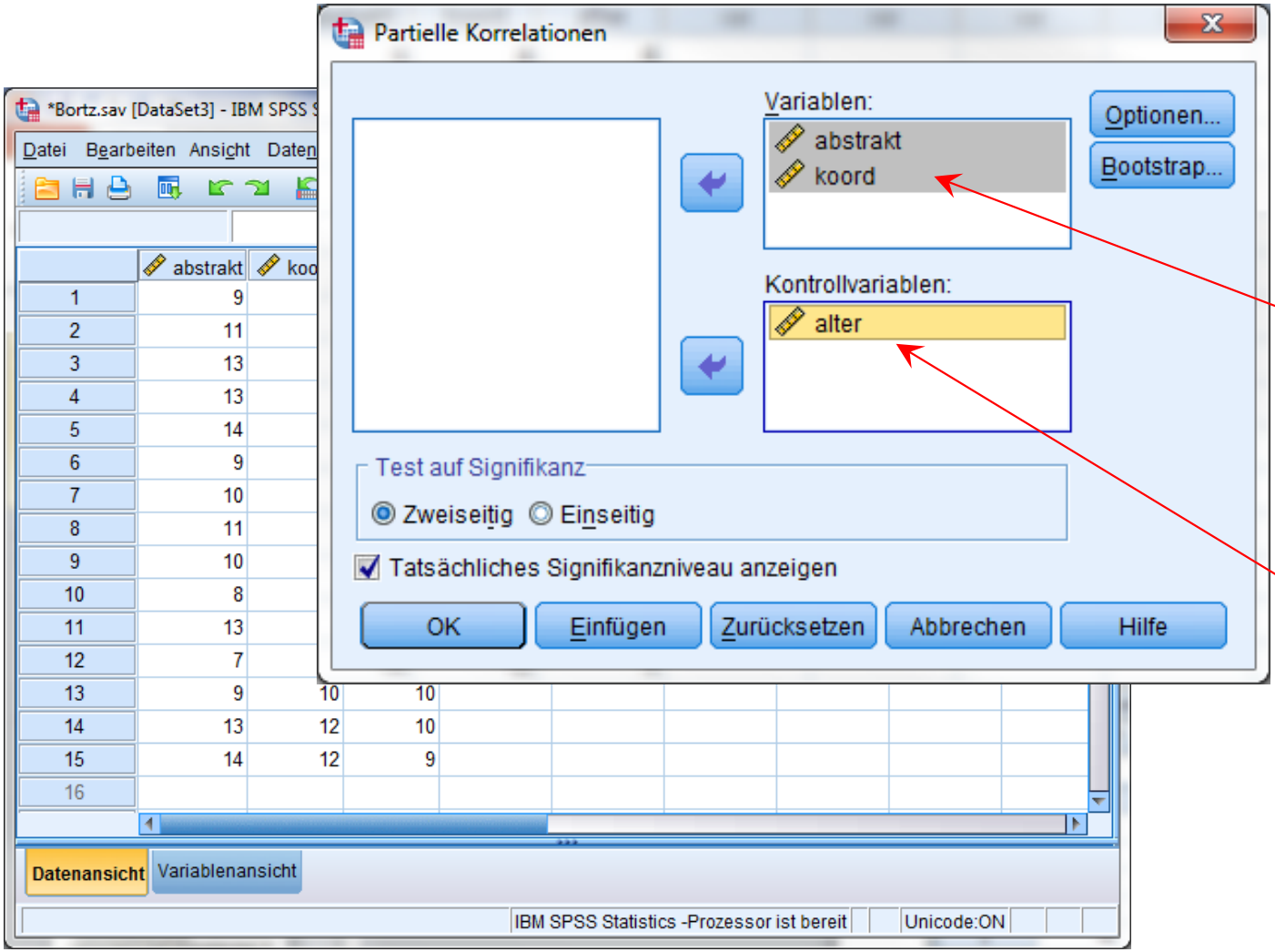
**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.



$r(\texttt{pers6}, \texttt{pers7}) = -.24$

Partialkorrelation

- In SPSS können Partialkorrelationen unter `Analysieren/Korrelation/Partiell…` angefordert werden (Prozedur: *PARTIAL CORR*):



Die Variablen X und Y sind unter „Variablen“ einzugeben.

Die Kontrollvariable(n) Z sind unter „Kontrollvariablen“ anzugeben.

Partielle Korrelation

Korrelationen

Kontrollvariablen			abstrakt	koord
alter	abstrakt	Korrelation	1,000	,722
		Signifikanz (zweiseitig)	.	,004
		Freiheitsgrade	0	12
	koord	Korrelation	,722	1,000
		Signifikanz (zweiseitig)	,004	.
		Freiheitsgrade	12	0

$r_{XY.Z} = 0.722$

Weitere Korrelationen

- Die dargestellte Produkt-Moment Korrelation ist nur dann zur Bestimmung der Enge des Zusammenhangs zweier Variablen X und Y geeignet, wenn beide Variablen (mindestens) intervallskaliert sind. Weisen eine oder beide Variable ein niedrigeres Skalenniveau auf, so sind andere Korrelationskoeffizienten heranzuziehen.
- Aus der auf der folgenden Seite dargestellten Tabelle kann entnommen werden, bei welcher Kombination von Skalenniveau in X und Y welche Korrelationskoeffizienten in Frage kommen. (Welche der beiden Variablen dabei X und welche Y ist, spielt keine Rolle.)
- Beim Nominalskalenniveau werden in der Tabelle nur die Korrelationskoeffizienten für den Fall dargestellt, dass die Variable(n) dichotom sind (d.h. nur zwei Ausprägungen aufweisen). In diesem Fall kann man unterscheiden, ob die Variable
  - natürlich dichotom vorliegt, d.h. es existieren prinzipiell nur zwei Kategorien (wie z.B. beim Geschlecht) oder
  - künstlich dichotom vorliegt, d.h. eine kontinuierliche Variable wurde vergröbernd in zwei Kategorien eingeteilt = dichotomisiert (z.B. die Alter in ≤ 30 und > 30 Jahre)
- Grundsätzlich gilt, dass prinzipiell immer auch die Korrelationen zulässig sind, die ein schwächeres Skalenniveau erfordern (also z.B. kann Spearmans r_s auch berechnet werden, wenn beide Variablen intervallskaliert sind). Sie sind meist aber weniger informativ.

Weitere Korrelationen

- Es gibt eine Vielzahl von Korrelationskoeffizienten. Das wichtigste Unterscheidungsmerkmal besteht darin, welches Skalenniveau die beiden Variablen X und Y aufweisen:

		Skalenniveau von X			
		intervall	ordinal	künstlich dichotom	natürlich dichotom
Skalenniveau von Y	intervall	Produkt-Moment Korrelation r	u.a. Rang-korrelationen nach Spearman r_s oder Kendall τ	biseriale Korrelation r_b	punktbiseriale Korrelation r_{pb}
	ordinal			biseriale Rangkorrelation	
	nominal: künst-lich dichotom			tetrachorische Korrelation r_{tet}	φ - Koeffizient
	nominal: natür-lich dichotom				

- Für alle im Folgenden dargestellten Korrelation gilt (sofern nicht anders angegeben):
  - Sie schwanken zwischen -1 (perfekter negativer Zusammenhang) und $+1$ (perfekter positiver Zusammenhang).
  - Der Wert 0 zeigt einen fehlenden Zusammenhang an.

Rangkorrelation nach Spearman

- Die Rangkorrelation nach Spearman r_s ist geeignet, um den Zusammenhang zwischen zwei (mindestens) ordinalskalierten Variablen X und Y zu quantifizieren.

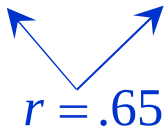
- Zur Bestimmung von r_s sind die Messwerte in beiden Variablen X und Y aufsteigend rangzuordnen: Jedem Messwert wird sein Rangplatz $\mathrm{Rg}()$ zugewiesen. Liegen Rangplatzbindungen (ties) vor, so werden mittlere Rangplätze vergeben.

- Die Rangkorrelation entspricht dann der Produkt-Moment-Korrelation zwischen den beiden Rangreihen $\mathrm{Rg}(X)$ und $\mathrm{Rg}(Y)$: $r_s = r[\mathrm{Rg}(X), \mathrm{Rg}(Y)]$.

- Beispiel: Es wird an 10 Schülern untersucht, wie stabil die Englischleistung in zwei aufeinanderfolgenden Klassenarbeiten ist. Es ist X = Punktzahl in der ersten Klassenarbeit und Y = Punktzahl in der zweiten Klassenarbeit.

  Die Rangkorrelation beträgt $r_s = 0.65$.

Vp	X	Y	$\mathrm{Rg}(X)$	$\mathrm{Rg}(Y)$
1	11	12	9	8
2	**3**	6	2.5	2
3	**3**	8	2.5	3
4	6	**10**	6	6
5	4	13	4	9
6	14	**10**	10	6
7	1	2	1	1
8	10	15	8	10
9	5	**10**	5	6
10	8	9	7	4



Rangkorrelation nach Spearman

- Liegen keine Ties vor, so lässt sich die Formel vereinfachend wie folgt schreiben:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}$$

mit d_i als Differenz der Rangplätze $\mathrm{Rg}(x_i) - \mathrm{Rg}(y_i)$.

- Wenden wir die obige Formel trotz bestehender Ties an, so erhalten wir:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot [(9-8)^2 + \ldots + (7-4)^2]}{10 \cdot (10^2 - 1)} = 0.66$$

- Im Beispiel ist die Abweichung aufgrund des geringen Tie-Anteils gering. Prinzipiell existieren für das Vorliegen von Ties Korrekturformeln (vgl. Bortz und Schuster, 2010, S. 179).

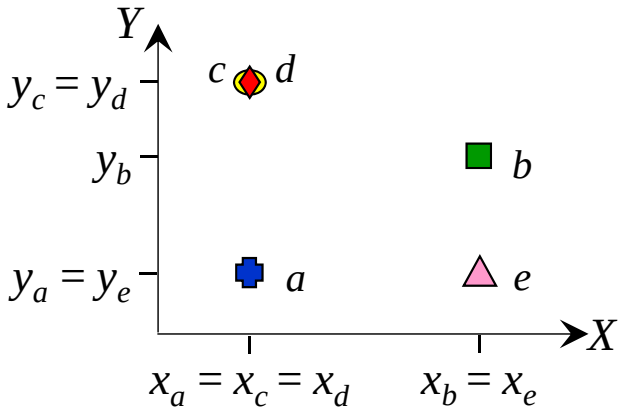
Vp	X	Y	$\mathrm{Rg}(X)$	$\mathrm{Rg}(Y)$
1	11	12	9	8
2	3	6	2.5	2
3	3	8	2.5	3
4	6	10	6	6
5	4	13	4	9
6	14	10	10	6
7	1	2	1	1
8	10	15	8	10
9	5	10	5	6
10	8	9	7	4

Rangkorrelationen

- Neben Spearmans Maß gibt es eine Familie von Koeffizienten, mit der ebenfalls die Stärke des Zusammenhangs zwischen zwei ordinalskalierten Variablen quantifiziert werden kann.
- Sie basieren auf einer anderen Logik und betrachten immer alle korrespondierenden Paare von Messwerten in X und Y (vollständiger Paarvergleich). Ein Messwertpaar (i, j) heißt ...

Bezeichnung	wenn gilt ...	Beispiel	Häufigkeit
konkordant	$x_i < x_j$ und $y_i < y_j$ oder $x_i > x_j$ und $y_i > y_j$	(a, b)	C
disko(nko)rdant	$x_i < x_j$ und $y_i > y_j$ oder $x_i > x_j$ und $y_i < y_j$	(c, e)	D
getied nur in X	$x_i = x_j$ und $y_i \neq y_j$	(a, d)	T_X
getied nur in Y	$y_i = y_j$ und $x_i \neq x_j$	(a, e)	T_Y
getied in X und Y	$x_i = x_j$ und $y_i = y_j$	(c, d)	T_{XY}

- Zur Verdeutlichung dient das (Mini-) Streudiagramm mit 5 Punkten a bis e rechts.
- Zur Berechnung der Koeffizienten müssen die Häufigkeiten der verschiedenen Typen von Messwertpaar-Relationen im Datensatz bestimmt werden (siehe Spalte oben rechts).



Rangkorrelationen

Vp	X	Y
1	6	4
2	7	8
3	6	7
4	9	4
5	7	8

$$\frac{n\cdot(n-1)}{2}=\frac{5\cdot(5-1)}{2}=10 \text{ Paare}$$
$$=C+D+T_X+T_Y+T_{XY}$$

	1	2	3	4	5
1		(1,2)	(1,3)	(1,4)	(1,5)
2			(2,3)	(2,4)	(2,5)
3				(3,4)	(3,5)
4					(4,5)
5					

Beispiel: In einem Wettkampf bewerten zwei Punktrichter die Ausführung der Übung (B-Note) von $n=5$ Bodenturnern auf einer Skala von 1 bis 10 (=Bestnote). Wie hoch ist der Zusammenhang der beiden Beurteilungsreihen?

Nr.	Paar	kon-kordant	diskor-dant	nur Tie in X	nur Tie in Y	Tie in X und Y
1	(1,2)	×				
2	(1,3)			×		
3	(1,4)				×	
4	(1,5)	×				
5	(2,3)	×				
6	(2,4)		×			
7	(2,5)					×
8	(3,4)		×			
9	(3,5)	×				
10	(4,5)		×			
		$C=4$	$D=3$	$T_X=1$	$T_Y=1$	$T_{XY}=1$

Rangkorrelationen

- Ein positiver Zusammenhang zwischen beiden Variablen liegt vor, wenn es mehr konkordante als diskordante Paare gibt. Ein Maß für die Enge des Zusammenhangs wäre daher $C - D$. Der Nachteil dieses Maßes liegt darin, dass es abhängig von der Stichprobengröße ist.

- Eine naheliegende Normierung an der Gesamtzahl an Paaren wurde von Kendall vorgeschlagen und nach ihm als Kendalls τ (Tau) benannt:

$$\tau = \tau_a = \frac{C - D}{n \cdot (n-1)/2} = \frac{C - D}{C + D + T_X + T_Y + T_{XY}} = \frac{2 \cdot (C - D)}{n \cdot (n-1)}$$

- Goodman und Kruskal haben ein Maß γ (Gamma) vorgeschlagen, bei dem die Normierung an der Summe der konkordanten und diskordanten Paare erfolgt:

$$\gamma = \frac{C - D}{C + D}$$

- Alternativ hat Kendall ein Maß vorgeschlagen, bei dem an dem geometrischen Mittel aus der Häufigkeit, mit der kein Tie in Y vorliegt und der Häufigkeit, mit der kein Tie in X vorliegt normiert wird:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}}$$

- Daneben gibt es noch eine Vielzahl weiterer Normierungsmöglichkeiten und Maße (auch asymmetrische, die nur Ties in einer der beiden Variablen berücksichtigen).

Rangkorrelationen

- Die Bestimmung anhand des Beispieldatensatzes ergibt:

Vp	X	Y
1	6	4
2	7	8
3	6	7
4	9	4
5	7	8

$$\tau = \frac{2 \cdot (C - D)}{n \cdot (n-1)} = \frac{2 \cdot (4-3)}{5 \cdot (5-1)} = 0.1$$

$$\gamma = \frac{C - D}{C + D} = \frac{4-3}{4+3} = 0.143$$

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}} = \frac{4-3}{\sqrt{(4+3+1) \cdot (4+3+1)}} = 0.125$$

$$r_s = .083$$

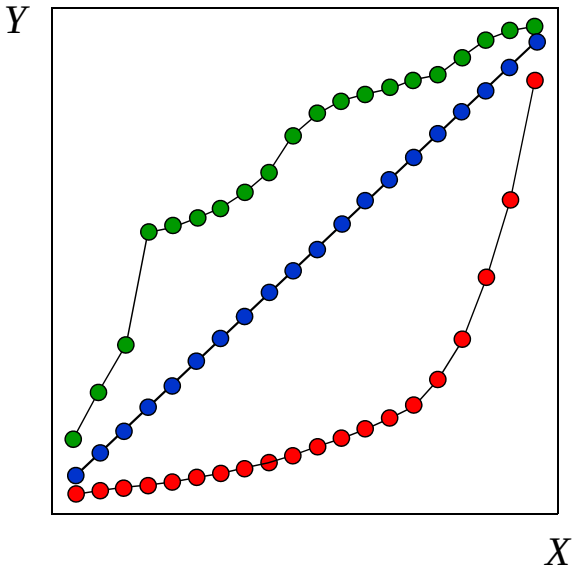
$C = 4, D = 3$

$T_X = T_Y = T_{XY} = 1$

Rangkorrelationen

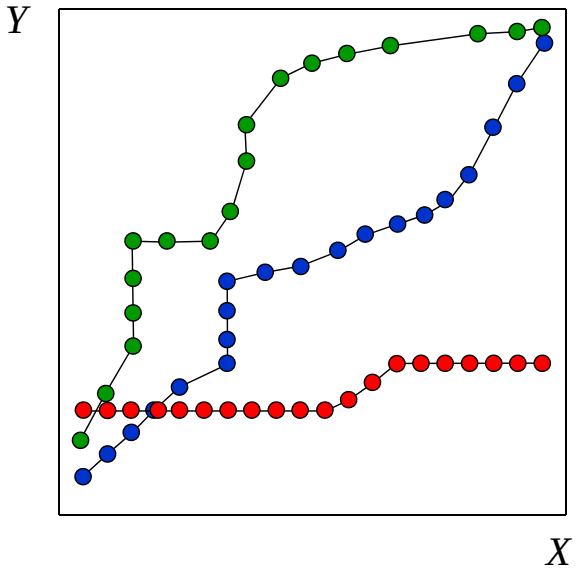
- Übereinstimmende Eigenschaften der drei Maße:
  - Sind alle Paare konkordant, ist also $C = n \cdot (n-1)/2$, so sind alle Koeffizienten gleich $+1$ (perfekter positiver Zusammenhang).
  - Sind alle Paare diskordant, ist also $D = n \cdot (n-1)/2$, so sind alle Koeffizienten gleich -1 (perfekter negativer Zusammenhang).
  - Gibt es gleich viele konkordante wie diskordante Paare, so ist der Koeffizient 0 (kein Zusammenhang).
  - Es gilt: $|\tau| \leq |\tau_b| \leq |\gamma|$. Existieren keine Ties, so sind die Koeffizienten gleich: $\tau = \tau_b = \gamma$.
  - Die Koeffizienten (auch r_s) sind invariant gegenüber streng monoton wachsenden Transformationen von X und Y, d.h. z.B. für X gegenüber allen Transformationen f, für die gilt: wenn $x_i < x_j$ dann ist auch $f(x_i) < f(x_j)$ (und natürlich auch linearen Transformationen).
  - Die Koeffizienten unterscheiden sich darin, wie sie Ties gewichten. Liegt ein hoher Tie-Anteil in den Variablen vor, so wirkt sich dies am stärksten mindernd auf τ, weniger auf τ_b und gar nicht systematisch auf γ aus.
  - Es erscheint daher sinnvoll, dann, wenn Ties durch die Prozedur der Datenerhebung erzwungen werden (z.B. bei einer dreistufigen Antwortskala), diese nicht als mindernd zu werten (also eher γ zu wählen).

Rangkorrelationen



- keine Ties

- Alle drei Kurven sind Beispiele für perfekte streng monoton steigende Zusammenhänge: Es gilt immer: Wenn $x_i < x_j$ dann $y_i < y_j$. (Analog mit „>“-Relation in Y bei monoton fallenden Zusammenhängen)
- Es existieren keine Ties.
- Es gilt: $\tau = \tau_b = \gamma = r_s = 1$.



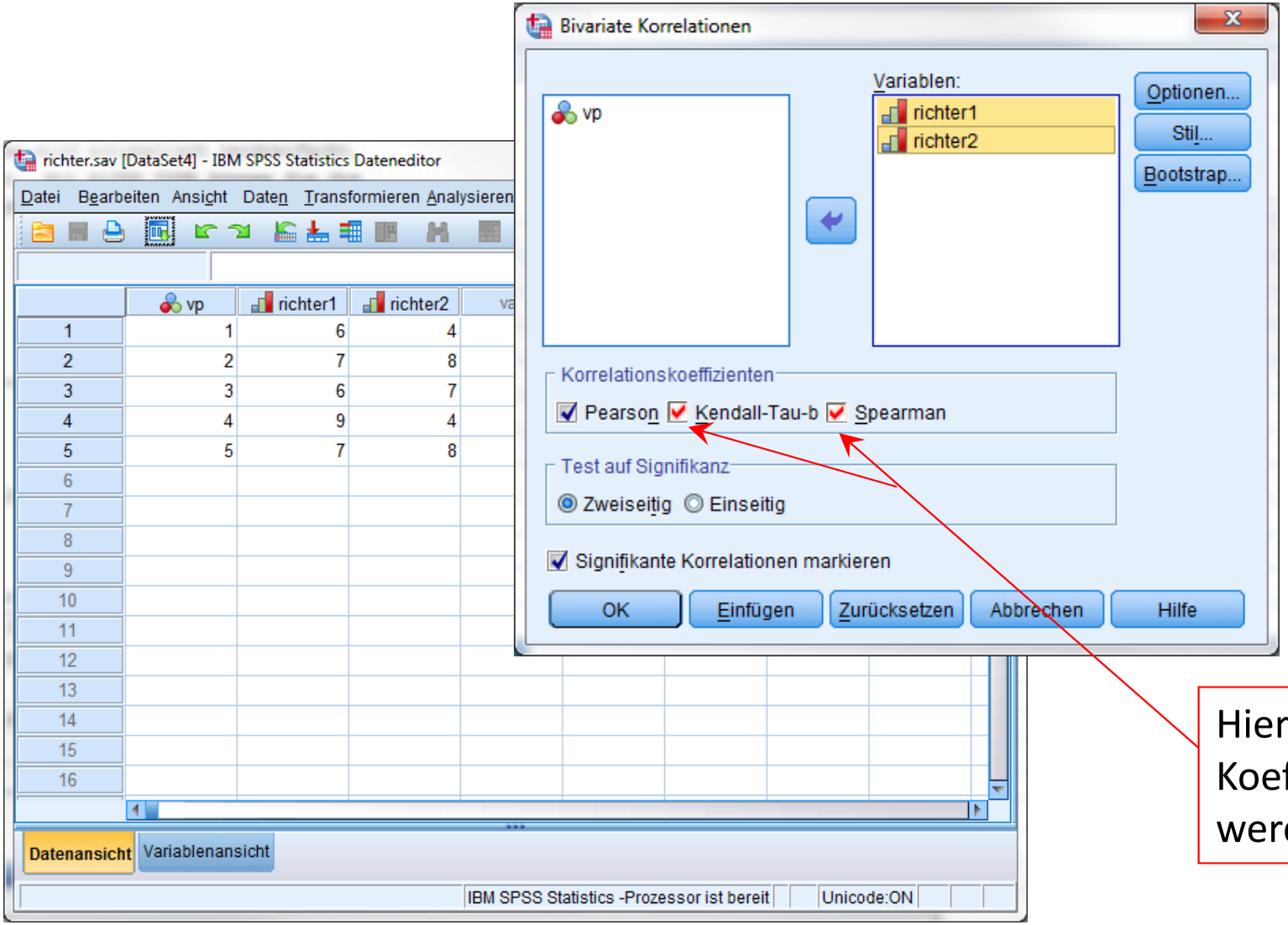
- nur Ties in Y ($T_Y > 0$)
- nur Ties in X ($T_X > 0$)
- nur Ties in X und nur Ties in Y ($T_X > 0$, $T_Y > 0$)

(Ties in X und Y zeigen sich in übereinanderliegenden Punkten)

- Alle drei Kurven sind Beispiele für perfekte schwach monoton steigende Zusammenhänge: Es gilt immer: Wenn $x_i < x_j$ dann $y_i \leq y_j$. (Analog mit „≥“-Relation in Y bei monoton fallenden Zusammenhängen)
- Es existieren Ties.
- Es gilt: $\gamma = 1$. τ, τ_b, $r_s < 1$.

Rangkorrelationen in SPSS

- Die Koeffizienten r_s und τ_b erhält man unter `Analysieren/Korrelation/Bivariat…`:



Hier können die beiden Koeffizienten aktiviert werden.

Rangkorrelationen in SPSS

Nichtparametrische Korrelationen

Korrelationen

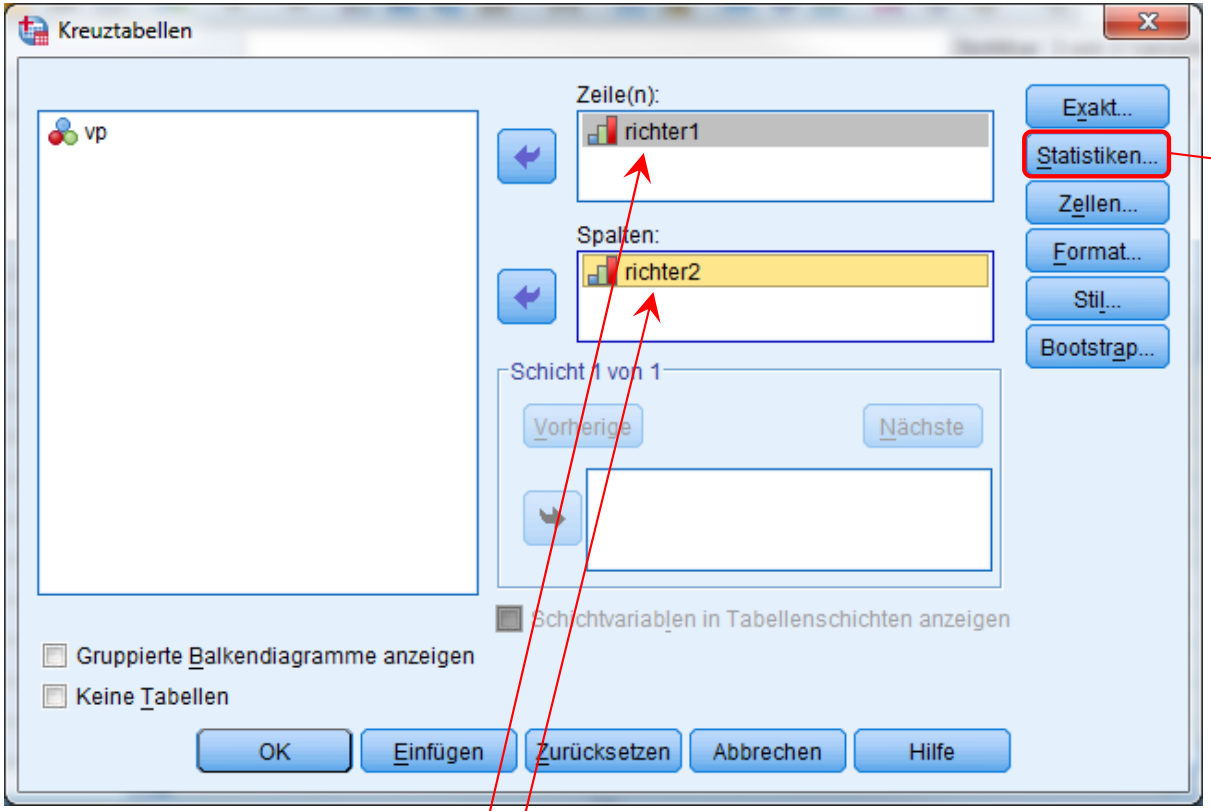
			richter1	richter2
Kendall-Tau-b	richter1	Korrelationskoeffizient	1,000	,125
		Sig. (2-seitig)	.	,782
		N	5	5
	richter2	Korrelationskoeffizient	,125	1,000
		Sig. (2-seitig)	,782	.
		N	5	5
Spearman-Rho	richter1	Korrelationskoeffizient	1,000	,083
		Sig. (2-seitig)	.	,894
		N	5	5
	richter2	Korrelationskoeffizient	,083	1,000
		Sig. (2-seitig)	,894	.
		N	5	5

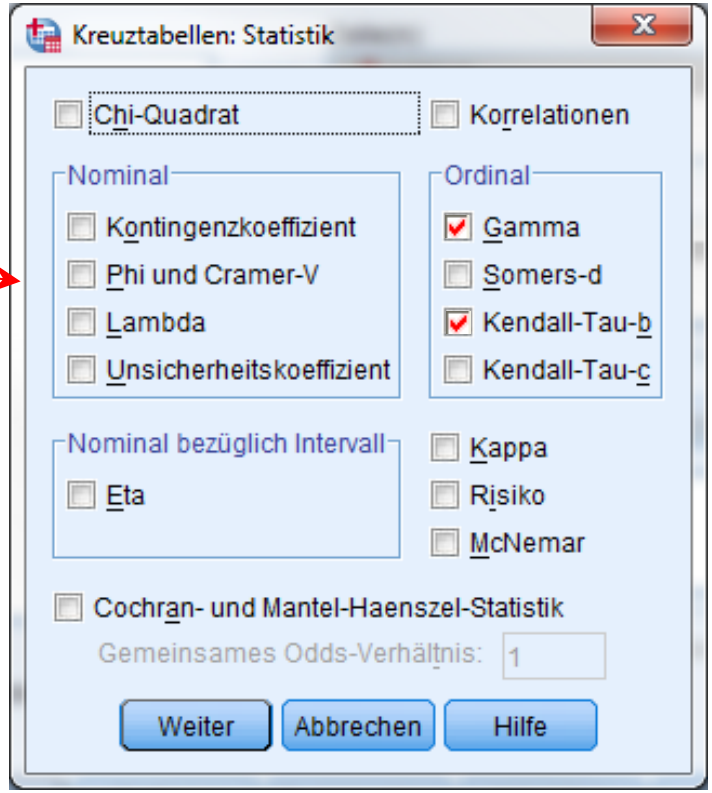
$\tau_b = .125$

$r_s = .083$

Rangkorrelationen in SPSS

- Die Koeffizienten γ und τ_b erhält man unter `Analysieren/Deskriptive_Statistiken/Kreuztabellen…`:





Hier muss eine Variable unter „Zeile(n)“ und eine unter „Spalten“ angegeben werden.

Im Dialog (`Statistiken`) können dann unter „Ordinal“ die beiden Statistiken angewählt werden.

Kreuztabellen

... (Tabellen weggelassen) ...

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises t[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-b	,125	,493	,256	,798
	Gamma	,143	,566	,256	,798
Anzahl der gültigen Fälle		5			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

$\gamma = .143$

$\tau_b = .125$

Weitere Korrelationen

		Skalenniveau von X			
		intervall	ordinal	künstlich dichotom	natürlich dichotom
Skalenniveau von Y	intervall	Produkt-Moment Korrelation r	u.a. Rang-korrelationen nach Spearman r_s oder Kendall τ	biseriale Korrelation r_b	punktbiseriale Korrelation r_{pb}
	ordinal			biseriale Rangkorrelation	
	nominal: künst-lich dichotom			tetrachorische Korrelation r_{tet}	φ - Koeffizient
	nominal: natür-lich dichotom				

Korrelation

- Mittels der punktbiserialen Korrelation r_{pb} kann man den Zusammenhang zwischen einer intervallskalierten Variablen X und einer natürlich dichotomen Variablen Y quantifizieren.

- Die punktbiseriale Korrelation entspricht der Produkt-Moment Korrelation zwischen X und Y: $r_{pb} = r(X, Y)$. Die Formel lässt sich wie folgt „vereinfachen“:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_X} \cdot \sqrt{\frac{n_1 \cdot n_0}{n \cdot (n-1)}}$$

mit $\bar{x}_1$ und $\bar{x}_0$ als Mittelwerte der beiden (z.B.) 0-1-codierten Gruppen und n_0 und n_1 als Zahl der Personen in diesen Gruppen.

- Beispiel: Bei 8 Mitarbeitern eines Unternehmens wurde die Arbeitszufriedenheit (X, Werte zwischen 0 und 30) sowie die Beobachtung, ob sie im darauffolgenden Jahr ihren Job gekündigt haben (Y, codiert 0=nein, 1=ja) erfasst.

  Es ergibt sich $r_{pb} = r = -0.24$. Wir haben die Kündigung als den höheren Wert codiert. Entsprechend besagt die Korrelation, dass die Mitarbeiter, die gekündigt haben, weniger arbeitszufrieden waren.

Vp	X	Y
1	29	0
2	18	1
3	22	1
4	19	0
5	12	0
6	24	0
7	27	0
8	19	1

$r = -0.24$

- Vorbereitende Berechnungen und die von r_{pb} mit der „vereinfachten“ Formel ergeben:

$$\bar{x}_0 = \frac{1}{5}(29+19+12+24+27) = 22.2 \qquad n = 8, n_1 = 3, n_0 = 5$$

$$\bar{x}_1 = \frac{1}{3}(18+22+19) = 19.667$$

$$\bar{x} = \frac{1}{8}(29+18+22+\ldots+19) = 21.25$$

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{(29-21.25)^2+\ldots+(19-21.25)^2}{8-1}} = 5.445$$

$$r_{pb} = \frac{\bar{x}_1-\bar{x}_0}{s_X}\cdot\sqrt{\frac{n_1\cdot n_0}{n\cdot(n-1)}} = \frac{19.667-22.2}{5.445}\cdot\sqrt{\frac{3\cdot 5}{8\cdot(8-1)}} = -0.24$$

Vp	X	Y
1	29	0
2	18	1
3	22	1
4	19	0
5	12	0
6	24	0
7	27	0
8	19	1

X_0 ($Y = 0$)
29
19
12
24
27

X_1 ($Y = 1$)
18
22
19

Korrelation

- Der Phi-Koeffizient φ (oder ϕ) quantifiziert den Zusammenhang zwischen zwei natürlich dichotomen Variablen X und Y.
- In diesem Fall gibt es nur vier verschiedene Kombinationen von Merkmalsausprägungen, deren Auftretenshäufigkeiten $n = a + b + c + d$ in einer Vierfeldertafel dargestellt werden können:

		Y		
		y_1	y_2	
X	x_1	a	b	$a + b$
	x_2	c	d	$c + d$
		$a + c$	$b + d$	n

Randhäufigkeiten

- Beispiel: Es soll untersucht werden, ob es einen Zusammenhang zwischen dem Geschlecht (X, codiert 0 = ♂, 1 = ♀) und dem Glauben an Horoskope (Y, codiert 0 = nein, 1 = ja) gibt.

Glauben an Horoskope

	0 = nein	1 = ja	
0 = ♂	5	2	7
1 = ♀	4	4	8
	9	6	15

Vp-Nr.	X	Y
1	0	0
2	1	0
3	1	1
4	0	1
5	1	1
6	1	0
7	0	0
8	1	0
9	0	0
10	1	1
11	0	1
12	0	0
13	1	1
14	1	0
15	0	0

Korrelation

- Der Phi-Koeffizient berechnet sich dann wie folgt (alle Randhäufigkeiten > 0):

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

	y_1	y_2	
x_1	a	b	$a+b$
x_2	c	d	$c+d$
	$a+c$	$b+d$	n

- Beispiel: Einsetzen der Häufigkeiten ergibt:

$$\varphi = \frac{5 \cdot 4 - 2 \cdot 4}{\sqrt{7 \cdot 8 \cdot 9 \cdot 6}} = 0.22$$

Es besteht also ein positiver Zusammenhang: Frauen glauben häufiger an Horoskope. (Bei der Interpretation des Vorzeichens muss man die Richtung der Codierung beachten. Eine Vertauschung der Codierung in einer Variable führt zur Umkehrung des Vorzeichens von φ.)

Glauben an Horoskope

	0 = nein	1 = ja	
♂ = 0	5	2	7
♀ = 1	4	4	8
	9	6	15

- Die φ-Korrelation ist identisch mit der Produkt-Moment-Korrelation zwischen den Messwertreihen: $\varphi = r$.

X	Y
0	0
1	0
1	1
0	1
1	1
1	0
0	0
1	0
0	0
1	1
0	1
0	0
1	1
1	0
0	0

$r = .22$

Korrelation

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

		Y: y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
		$a+c$	$b+d$	n

- Der φ-Koeffizient kann nur dann die Werte –1 und +1 annehmen, wenn eine der beiden Diagonalen a, d oder b, c nur Häufigkeiten von 0 enthält.
- Dies ist nur möglich, wenn beide Randverteilungen gleich sind.

	0 = nein	1 = ja	
♂ = 0	7	0	7
♀ = 1	0	8	8
	7	8	15

- Frage: Was wäre der maximal mögliche positive Zusammenhang im Beispieldatensatz?

Einsetzen erbringt: $\frac{7 \cdot 6 - 0 \cdot 2}{\sqrt{7 \cdot 8 \cdot 9 \cdot 6}} = 0.764$

- Für gegebene Randverteilungen lassen sich das maximal und minimal mögliche φ allgemein wie folgt bestimmen:

$$\varphi_{max} = \min\left(\sqrt{\frac{(a+b) \cdot (b+d)}{(a+c) \cdot (c+d)}}, \sqrt{\frac{(a+c) \cdot (c+d)}{(a+b) \cdot (b+d)}} \right)$$

$$\varphi_{min} = \max\left(-\sqrt{\frac{(a+b) \cdot (a+c)}{(c+d) \cdot (b+d)}}, -\sqrt{\frac{(c+d) \cdot (b+d)}{(a+b) \cdot (a+c)}} \right)$$

	0 = nein	1 = ja	
♂ = 0	5	2	7
♀ = 1	4	4	8
	9	6	15

	0 = nein	1 = ja	
♂ = 0	7	0	7
♀ = 1	2	6	8
	9	6	15

- Im Beispiel ergibt sich:

$$\varphi_{max} = \min\left(\sqrt{\frac{(a+b)\cdot(b+d)}{(a+c)\cdot(c+d)}}, \sqrt{\frac{(a+c)\cdot(c+d)}{(a+b)\cdot(b+d)}}\right)$$

$$= \min\left(\sqrt{\frac{7\cdot 6}{9\cdot 8}}, \sqrt{\frac{9\cdot 8}{7\cdot 6}}\right) = \min(0.764, 1.309) = 0.764$$

$$\varphi_{min} = \max\left(-\sqrt{\frac{(a+b)\cdot(a+c)}{(c+d)\cdot(b+d)}}, -\sqrt{\frac{(c+d)\cdot(b+d)}{(a+b)\cdot(a+c)}}\right)$$

$$= \max\left(-\sqrt{\frac{7\cdot 9}{8\cdot 6}}, -\sqrt{\frac{8\cdot 6}{7\cdot 9}}\right) = \max(-1.146, -0.873) = -0.873$$

		Y		
		y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
		$a+c$	$b+d$	n

	0 = nein	1 = ja	
♂ = 0	5	2	7
♀ = 1	4	4	8
	9	6	15

- φ kann also bei diesen Randverteilungen nur Werte im Intervall $[-0.873, 0.764]$ annehmen. Der beobachtete Wert von $\varphi = 0.22$ liegt beträchtlich unter dem maximal möglichen Wert.

- Hinweis: Hätten wir wie nebenstehend eine der beiden Variablen anders herum codiert, so hätte sich entsprechend ein $\varphi = -0.22$ ergeben und das Intervall würde lauten $[-0.764, 0.873]$.

	0 = ja	1 = nein	
♂ = 0	2	5	7
♀ = 1	4	4	8
	6	9	15

Korrelation

- Zur Betrachtung der Frage, ob zwischen zwei dichotomen Variablen ein Zusammenhang besteht, kann man sich auch ansehen, ob sich die relativen Anteile einer Merkmalausprägung in X in den beiden Y-Gruppen unterscheiden (oder umgekehrt). Wenn dies der Fall ist, besteht ein Zusammenhang und es besteht eine Abhängigkeit zwischen beiden Variablen. Andernfalls besteht kein Zusammenhang und beide Variablen sind unabhängig voneinander.

- Beispiel: An einer Stichprobe von 100 Personen wurden die beiden dichotomen Variablen Geschlecht (X, codiert $0 = ♀$, $1 = ♂$) und Rauchverhalten erhoben (Y, codiert $0 =$ Nichtraucher, $1 =$ Raucher).

- Besteht hier ein Zusammenhang? Man kann erkennen, dass der Anteil der Raucher bei den Frauen mit

  $13 / 50 = 0.26$

  niedriger ist als bei den Männern mit

  $17 / 50 = 0.34$

  (also 26% vs. 34%). Es besteht also ein Zusammenhang.

	0 = Nichtraucher	1 = Raucher	
0 = ♀	37	13	50
1 = ♂	33	17	50
	70	30	100

Korrelation

- Wie sähen die Häufigkeiten aus, wenn kein Zusammenhang zwischen X und Y bestehen würde? Um dies näher zu betrachten notieren wir im Folgenden die Häufigkeiten in der Vierfeldertafel alternativ wie folgt:

		Y: y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
		$a+c$	$b+d$	n

	Y: y_1	y_2	
x_1	h_{11}	h_{12}	$h_{1\bullet}$
x_2	h_{21}	h_{22}	$h_{2\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$	n

	Y: y_1	y_2	
x_1	f_{11}	f_{12}	$f_{1\bullet}$
x_2	f_{21}	f_{22}	$f_{2\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	1

- Es ist also z.B. $b = h_{12}$ die Häufigkeit, mit der Personen gleichzeitig in der Variable X den Wert x_1 und in der Variablen Y den Wert y_2 aufweisen (im Beispiel also die Zahl der weiblichen Raucher 13). Bei h_{ij} steht erste Subskript immer für die Zeile i und das zweite für die Spalte j.
- Summiert man die Häufigkeiten über beide Zeilen bzw. beide Spalten, so erhält man die vier Randhäufigkeiten, z.B. $b + d = h_{12} + h_{22} = h_{\bullet 2}$ (im Beispiel rechts die Zahl der Raucher = 30)
- In der Vierfeldertafel ganz rechts werden dann relative Häufigkeiten dargestellt: $f_{ij} = h_{ij} / n$.

	Nicht-raucher	Raucher	
♀	37	13	50
♂	33	17	50
	70	30	100

Korrelation

- Wie sähen die Häufigkeiten in den Zellen aus, wenn bei gleichen Randverteilungen kein Zusammenhang zwischen X und Y (Unabhängigkeit) existieren würde?
- Hier sollte der Anteil der Raucher bei Männern und Frauen gleich sein (oder umgekehrt der Anteil der Männer bei Rauchern und Nichtrauchern gleich sein). Dies ist im Beispiel relativ einfach durch Probieren herauszufinden.
- Dann gilt: Der Anteil der Raucher ist bei Männern und Frauen gleich: $15/50 = 0.3$ (also 30%); entsprechend der Anteil der Nichtraucher natürlich auch (70%).
- Und: Der Anteil der Männer ist bei Rauchern und Nichtrauchern gleich: $15/30 = 35/70 = 0.5$ (also 50%); entsprechend der Anteil der Frauen auch (50%).

	Nicht-raucher	Raucher	
♀	?	?	50
♂	?	?	50
	70	30	100

	Nicht-raucher	Raucher	
♀	35	15	50
♂	35	15	50
	70	30	100

- Wie kann man bei Unabhängigkeit der beiden Variablen die zu erwartenden Häufigkeiten allgemein bestimmen?
- Unter Verwendung des Multiplikationstheorems ergibt sich die erwartete relative Häufigkeit als Produkt der korrespondierenden relativen Randhäufigkeiten $f_{i\bullet} \cdot f_{\bullet j}$ bzw. für die erwartete absolute Häufigkeit e_{ij} in der Zelle (i, j) der Vierfeldertafel (wg. $h_{ij} = f_{ij} \cdot n$):

$$e_{ij} = (f_{i\bullet} \cdot f_{\bullet j}) \cdot n = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$$

Das Multiplikationstheorem der Wahrscheinlichkeitstheorie besagt, dass die Wahrscheinlichkeit des gemeinsamen Auftretens zweier Ereignisse (z.B. die Wahrscheinlichkeit zwei 6en zu würfeln) im Falle unabhängiger Ereignisse sich aus dem Produkt der Auftretenswahrscheinlichkeiten beider einzelnen Ereignisse ergibt (im Beispiel also 1/6 · 1/6 = 1/36).

Im Beispiel:

h_{ij}	Nichtraucher	Raucher	
♀	37	13	50
♂	33	17	50
	70	30	100

$$e_{11} = \frac{h_{1\bullet} \cdot h_{\bullet 1}}{n} = \frac{50 \cdot 70}{100} = 35$$

$$e_{12} = \frac{h_{1\bullet} \cdot h_{\bullet 2}}{n} = \frac{50 \cdot 30}{100} = 15$$

...

e_{ij}	Nichtraucher	Raucher
♀	35	15
♂	35	15

- Die Abweichungen der beobachteten von den bei Unabhängigkeit erwarteten Zellhäufigkeiten können wir nun quadrieren, an den erwarteten Häufigkeiten relativierten, und über alle Zellen summieren und erhalten folgende als χ^2-Quadrat (chi) bezeichnete Größe:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(h_{ij}-e_{ij})^2}{e_{ij}}$$

- Im Beispiel resultiert:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(h_{ij}-e_{ij})^2}{e_{ij}}$$

$$= \frac{(37-35)^2}{35} + \frac{(13-15)^2}{15} + \frac{(33-35)^2}{35} + \frac{(17-15)^2}{15}$$

$$= 0.762$$

h_{ij}

	Nichtraucher	Raucher	
♀	37	13	50
♂	33	17	50
	70	30	100

e_{ij}

	Nichtraucher	Raucher	
♀	35	15	50
♂	35	15	50
	70	30	100

- Weichen die empirisch beobachteten Häufigkeiten nicht von den Häufigkeiten unter Unabhängigkeit ab, so resultiert ein χ^2-Wert von 0, der also einen fehlenden Zusammenhang anzeigt. Je größer χ^2, desto enger ist der Zusammenhang und desto größer ist die Abhängigkeit zwischen den Variablen.

- Der χ^2-Wert steht in einem einfachen Zusammenhang mit dem φ-Koeffizient:

$$|\varphi| = \sqrt{\frac{\chi^2}{n}}$$

- Im Beispiel:

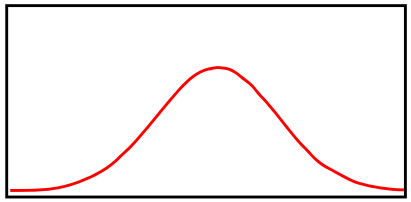
h_{ij}	Nicht-raucher	Raucher	
♀	37	13	50
♂	33	17	50
	70	30	100

$$|\varphi| = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{0.762}{100}} = .087$$

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}} = \frac{37 \cdot 17 - 13 \cdot 33}{\sqrt{50 \cdot 50 \cdot 70 \cdot 30}} = .087$$

Korrelation

- Für alle bisher dargestellten Korrelationsmaße, die den Zusammenhang unter Beteiligung mindestens einer natürlich kategorialen Variablen quantifizieren, gibt es Pendants, die zum Einsatz kommen können, wenn alle kategorialen Unterteilungen künstlich zustande gekommen sind:
  - Im Falle zweier dichotomen Variablen ist dies statt des φ-Koeffizienten die tetrachorische Korrelation r_{tet}.
  - Im Falle einer intervallskalierten mit einer dichotomen Variablen ist dies statt der punktbiserialen Korrelation r_{pb} die biseriale Korrelation r_b.
- Unter „künstlich“ wird dabei verstanden, dass davon ausgegangen wird, dass der nominalen Variable (latent) eine kontinuierliche Variable zugrunde liegt, die zudem normalverteilt ist.
- Es ist nicht immer klar, wann man von einer zugrundeliegenden kontinuierlichen Variable ausgehen kann. Z.B. bei einer Mathematikklausur (bestanden, nicht bestanden), Suizidversuch (ja, nein), Kündigung (ja, nein), Geschlecht (♀, ♂)?
- Auch die Annahme der Normalverteilung kann dabei verletzt sein (z.B. Alter, Suizidneigung). Ist dies der Fall, so ergeben die Koeffizienten verzerrte Ergebnisse.

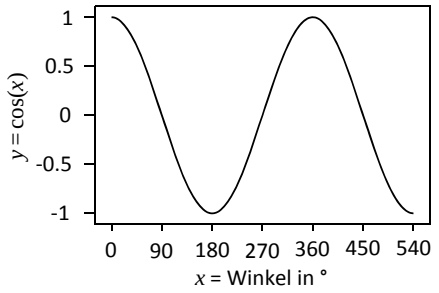


Korrelation

- Eine Näherungsformel für die tetrachorische Korrelation r_{tet} zwischen zwei künstlich dichotomen Variablen lautet (für $b, c > 0$):

$$r_{tet} = \cos\left(\frac{180°}{1+\sqrt{\frac{a \cdot d}{b \cdot c}}}\right)$$

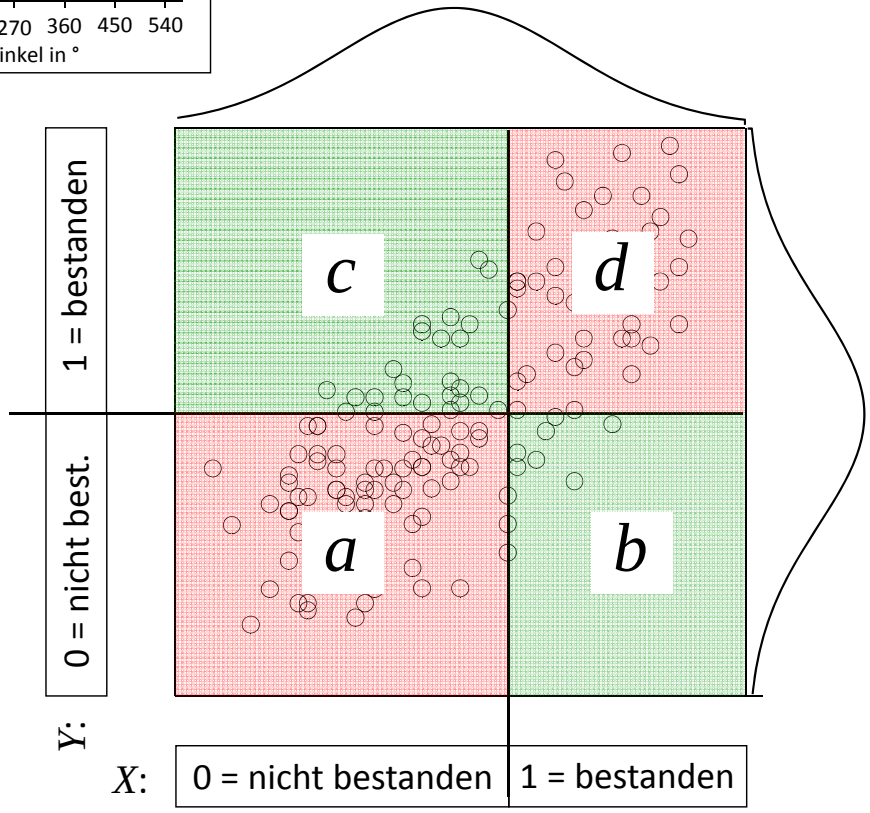
$$\cos(0°) = 1$$
$$\cos(90°) = 0$$
$$\cos(180°) = -1$$



- Je größer $a \cdot d$ relativ zu $b \cdot c$, desto größer das Kreuzproduktverhältnis $a \cdot d / (b \cdot c)$, desto größer der Nenner und desto mehr geht der Quotient gegen 0 und r_{tet} gegen 1.

- Beispiel: Besteht zwischen der Leistung in einer Mathematikklausur (X) und der Leistung in einer Biologieklausur (Y) ein Zusammenhang?

- Es wird davon ausgegangen, dass die Klausuren das kontinuierliche latente Kontinuum „Mathematik- bzw. Biologiekompetenz“ künstlich zerschneiden.



- Beispiel: Für die $n = 477$ (fiktive) Schüler ergab sich die untenstehende Vierfeldertafel.
- Einsetzen ergibt:

$$r_{tet} = \cos\left(\frac{180°}{1+\sqrt{\frac{a \cdot d}{b \cdot c}}}\right) = \cos\left(\frac{180°}{1+\sqrt{\frac{10 \cdot 404}{47 \cdot 16}}}\right)$$

$$= \cos\left(\frac{180°}{1+2.318}\right) = 0.584$$

		Y	
		y_1	y_2
X	x_1	c	d
	x_2	a	b

Biologieklausur

h_{ij}	0=nicht bestanden	1= bestanden	
Mathe-Klausur: 1 = bestanden	16	404	420
Mathe-Klausur: 0 = nicht bestanden	10	47	57
	26	451	477

- Zum Vergleich: $\varphi = 0.196$.
- (Bei sehr asymmetrischen Randverteilungen wie hier überschätzt obige Näherung r_{tet}. Eine genauere Schätzung, z.B. über einen SPSS-Macro von Enzmann (2007), ergibt $r_{tet} = 0.45$)

Weitere Korrelationen

		Skalenniveau von X			
		intervall	ordinal	künstlich dichotom	natürlich dichotom
Skalenniveau von Y	intervall	Produkt-Moment Korrelation r	u.a. Rang-korrelationen nach Spearman r_s oder Kendall τ	biseriale Korrelation r_b	punktbiseriale Korrelation r_{pb}
	ordinal			biseriale Rangkorrelation	
	nominal: künst-lich dichotom			tetrachorische Korrelation r_{tet}	φ-Koeffizient
	nominal: natür-lich dichotom				

- Für die beiden nicht behandelten Korrelationen finden Sie hier Angaben:
  - Biseriale Korrelation: Bortz & Schuster (2010, S. 172f)
  - Biseriale Rangkorrelation: Bortz & Schuster (2010, S. 177ff)
- Im Folgenden soll noch eine Korrelation dargestellt werden, die nicht in das obige Schema passt. Sie stellt die Erweiterung des φ-Koeffizienten dar, wenn eine Variable mehr als zwei Kategorien aufweist.

Korrelation

- Ist von zwei nominalskalierten Variablen mindestens eine nicht mehr dichotom sondern poly(cho)tom (d.h. weist mehr als zwei Ausprägungen auf), so wird aus der $2 \; x \; 2$ Vierfeldertafel eine $k \; x \; m$ Kontingenztafel, in der
  - h_{ij} die Auftretenshäufigkeit der i-ten Ausprägung in X und der j-ten Ausprägung in Y bezeichnet ($i = 1 \ldots k$ und $j = 1 \ldots m$).
  - Ein $\bullet$ für die Aggregierung über das betreffende Subskript dient wieder zur Bezeichnung der Randhäufigkeiten, also

$$h_{\bullet j} = \sum_{i=1}^{k} h_{ij}$$

$$h_{i\bullet} = \sum_{j=1}^{m} h_{ij}$$

  - Entsprechend gilt:

$$h_{\bullet\bullet} = \sum_{j=1}^{m} h_{\bullet j} = \sum_{i=1}^{k} h_{i\bullet} = n$$

X \ Y	y_1	y_2	…	y_j	…	y_m	
x_1	h_{11}	h_{12}		h_{1j}		h_{1m}	$h_{1\bullet}$
x_2	h_{21}	h_{22}		h_{2j}		h_{2m}	$h_{2\bullet}$
…							
x_i	h_{i1}	h_{i2}		h_{ij}		h_{im}	$h_{i\bullet}$
…							
x_k	h_{k1}	h_{k2}		h_{kj}		h_{km}	$h_{k\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$		$h_{\bullet j}$		$h_{\bullet m}$	$h_{\bullet\bullet}$

Korrelation

- Für die Quantifizierung des Zusammenhang zwischen zwei nominalen Variablen, von denen mindestens eine polytom ist, lässt sich Cramérs V heranziehen. (X weise dabei k und Y weise m Ausprägungen auf.)

- Dafür muss zunächst wieder die χ^2-Statistik errechnet werden, allgemein:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}}$$

- Auf dieser Basis ergibt sich dann:

$$V = \sqrt{\frac{\chi^2}{n \cdot [\mathrm{Min}(k, m) - 1]}}$$

	h_{ij}	Y: gute Bezahlung	nette Kollegen	interessante Arbeit	
X	Psychologie	8	19	23	50
	Jura	16	12	17	45
		24	31	40	95

- Fiktives Beispiel: Insgesamt 95 Studierende aus den Studienfächern Psychologie und Jura (X, codiert 1=Psychologie und 2=Jura) wurden danach gefragt, was ihnen bei der Wahl des künftigen Arbeitsplatzes am wichtigsten ist (Arbeitswerte Y, mit Optionen 1=„gute Bezahlung“, 2=„nette Kollegen“ und 3=„interessante Arbeit“). Es interessiert die Frage, ob beide Variablen voneinander unabhängig sind oder ein Zusammenhang besteht.

- Wir benötigen zunächst χ^2:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(h_{ij}-e_{ij})^2}{e_{ij}}$$

h_{ij}

	Bez.	Kollegen	int. Arbeit	
Psychologie	8	19	23	50
Jura	16	12	17	45
	24	31	40	95

$$e_{11} = 50 \cdot 24 / 95 = 12.632 \qquad \text{usw.}$$

$$\chi^2_{11} = \frac{(8-12.632)^2}{12.632} = 1.698 \qquad \text{usw.}$$

$$\chi^2 = 1.698 + 0.442 + \ldots + 0.200 = 4.898$$

e_{ij}

	Bez.	Kollegen	int. Arbeit
Psychologie	12.632	16.316	21.053
Jura	11.368	14.684	18.947

- Einsetzen in V führt dann zu:

$$V = \sqrt{\frac{\chi^2}{n \cdot [\mathrm{Min}(k,m)-1]}}$$

$$= \sqrt{\frac{4.898}{95 \cdot [\mathrm{Min}(2,3)-1]}} = 0.227$$

χ^2_{ij}

	Bez.	Kollegen	int. Arbeit
Psychologie	1.698	0.442	0.180
Jura	1.887	0.491	0.200

- Es gibt einen Zusammenhang zwischen dem Studienfach und den Arbeitswerten von $V = 0.23$.

- Die Eigenschaften des Koeffizient V sind:
  - Er liegt zwischen 0 (=kein Zusammenhang) und 1 (=perfekter Zusammenhang). Bei mehr als zwei nominalen Kategorien macht ein „je mehr X desto mehr/weniger Y“, bzw. eine Unterscheidung in einen positiven bzw. negativen Zusammenhang keinen Sinn mehr.
  - Der Koeffizient φ ist ein Spezialfall von V für $k = m = 2$: $V = |\varphi|$:

$$V = \sqrt{\frac{\chi^2}{n \cdot [\mathrm{Min}(k,m) - 1]}} \overset{k=m=2}{\Rightarrow} \sqrt{\frac{\chi^2}{n \cdot [\mathrm{Min}(2,2) - 1]}} = \sqrt{\frac{\chi^2}{n}} = |\varphi|$$

Korrelation in SPSS: Überblick

Korrelation	Prozedur in SPSS: Analysieren/...	Option
Produkt-Moment r	Korrelation/Bivariat	„Pearson“ (Weglasswert)
Rangkorrelation nach Spearman r_s	Korrelation/Bivariat	„Spearman“
Kendalls τ_b	Korrelation/Bivariat oder Deskriptive_Statistiken/ Kreuztabellen	„Kendall-Tau-b“ (Statistik): „Kendall-Tau-b“
Goodman & Kruskals γ	Deskriptive_Statistiken/ Kreuztabellen	„Gamma“
Punktbiseriale Korrelation r_{pb}	Korrelation/Bivariat	identisch mit r
φ-Koeffizient	Korrelation/Bivariat oder Deskriptive_Statistiken/ Kreuztabellen	identisch mit r (Statistik): „Phi und Cramer-V“
Cramérs V	Deskriptive_Statistiken/ Kreuztabellen	(Statistik): „Phi und Cramer-V“

- Folgende Korrelationen sind nicht in SPSS implementiert: τ, r_b, r_{tet}. Falls man sie braucht, (1) findet man im Internet Macros für SPSS oder (2) rechnet per Hand oder (3) verwendet ein anderes Statistik-Programm.

Produkt-Moment Korrelation

- Wann ist eine Korrelation hoch?
- Man kann Korrelationen quadrieren und erhält dann ein Maß für die „erklärte Varianz“.
- Eine weitere grobe Orientierung gibt Cohen (1988). Er bezeichnet Korrelation in der Größenordnung von
  - ≈ 0.10 als kleinen Effekt
  - ≈ 0.30 als mittleren Effekt
  - ≈ 0.50 als starken Effekt
- Die Bewertung einer Korrelation als hoch oder niedrig hängt aber vor allem stark vom Kontext ab (vgl. die beiden Beispiele auf der nächsten Folie).

Produkt-Moment Korrelation

- Beispiel 1: In einer groß angelegten über 5 Jahre dauernden Studie (vgl. Rosnow & Rosenthal, 1989) nahmen in einer Stichprobe von 22071 Ärzten etwa die Hälfte der Ärzte täglich eine Aspirin-Tablette und die andere Hälfte ein Placebo ein. Untersucht wurde, wie sich die Einnahme auf das Auftreten von Herzinfarkten auswirkt.

		Herzinfarkt	
		nein	ja
Medi-kation	Aspirin	10933	104
	Placebo	10845	189

$\varphi = 0.034$, $\varphi^2 = 0.001$, d.h. die Varianzaufklärung beträgt 0.1%.

Aber: 85 Ärzte weniger haben mit Aspirin einen Herzinfarkt erlitten (13 weniger tödlich).

- Beispiel 2: In einer Studie, in der untersucht wurde, wie man mit einem neuen Testverfahren zur Erfassung der sozialen Intelligenz den Ausbildungserfolg von Sozialarbeitern vorhersagen kann, ergibt sich eine Korrelation der Testergebnisse mit dem Ausbildungserfolg von 0.20. Aus der Literatur weiß man aber, dass man mit klassischen Intelligenztests, z.B. dem Intelligenz-Struktur-Test, bei Auszubildenden mit Zusammenhängen von $r = 0.32$ rechnen kann (Hülsheger, Maier, Stumpp, Maier & Muck, 2000).

Bivariate deskriptive Statistik: Zitierte Quellen

Zitierte Quellen

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Enzmann, D. (2002). SPSS-Macro r_bis, Version 2.1. Herunterladbar unter: http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html.
- Enzmann, D. (2007). SPSS-Macro TetCorr, Version 2.3. Herunterladbar unter: http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html.
- Hülsheger, U. R., Maier, G. W., Stumpp, T. & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland: Ergebnisse einer Metaanalyse. *Zeitschrift für Personalpsychologie*, *5*, 145-162.
- Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.