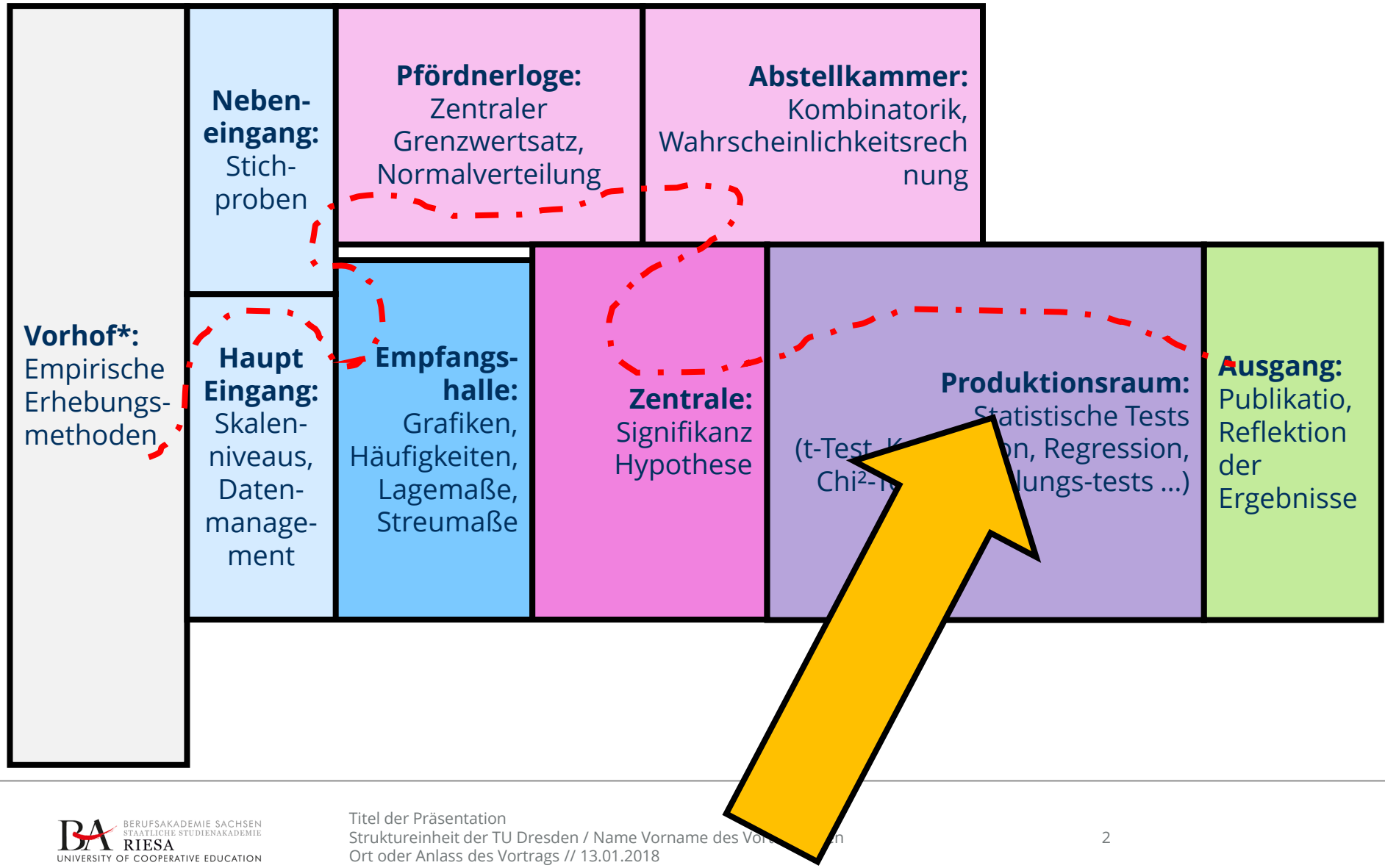
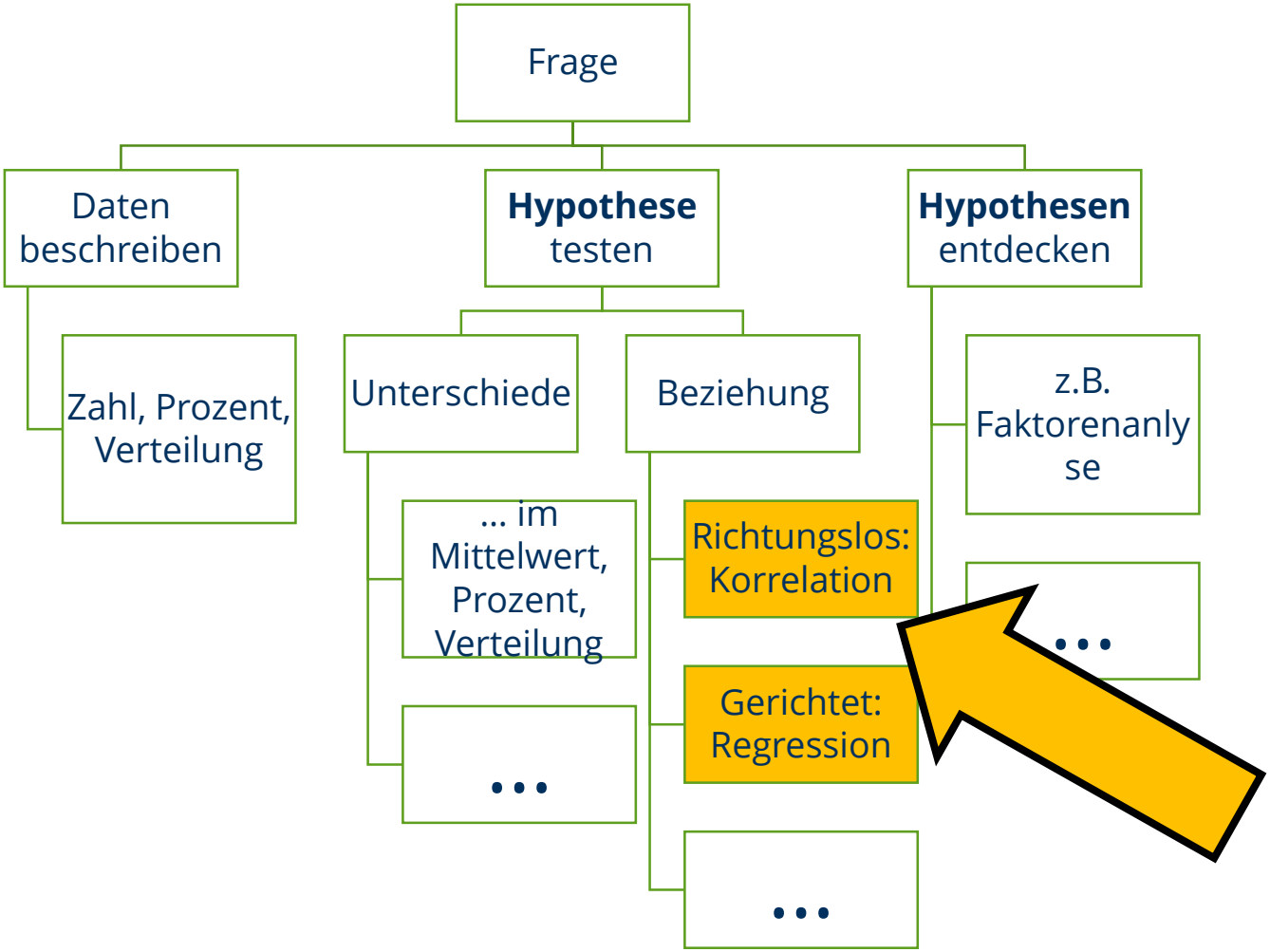
BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

Korrelation

Wo sind wir



Wo sind wir?



Korrelation: Idee und Streudiagramm

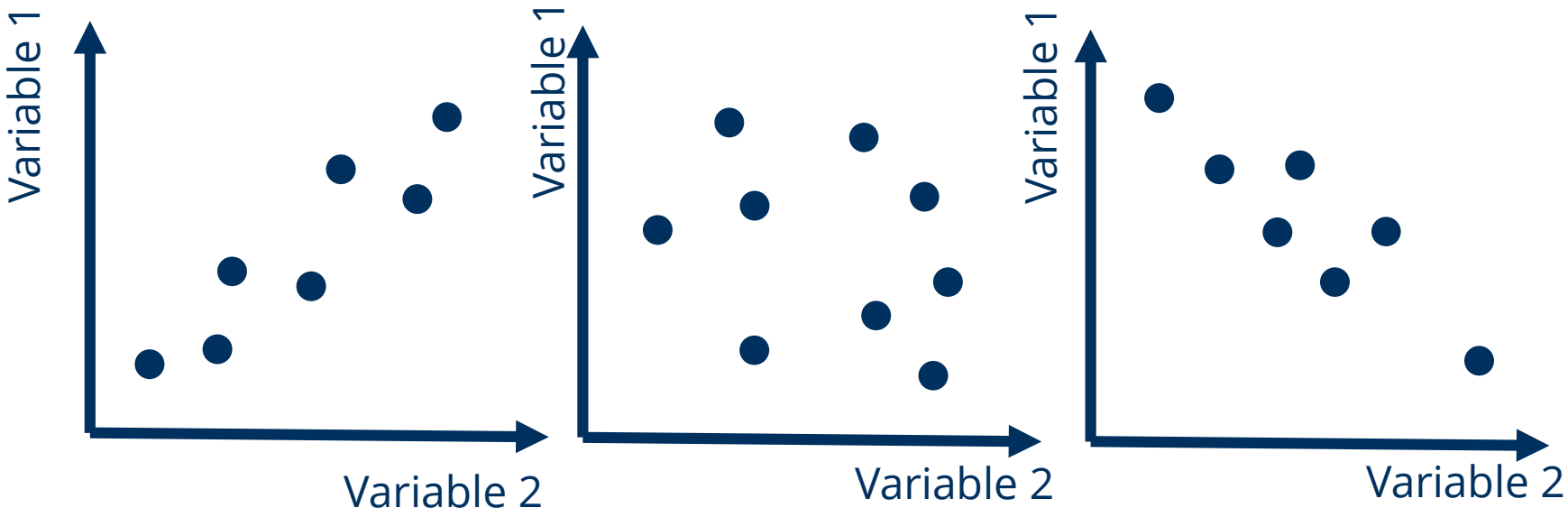
Die Korrelation beantwortet die Frage, gibt es einen Zusammenhang zwischen zwei Variablen.

Beispiel: Gibt es einen Zusammenhang zwischen Motivation und Arbeitsleistung?

Motivation	Arbeitsleistung
1	2
2	4
3	1
4	3
5	7
6	5
7	6
8	9
9	10
10	8

Das Streu- oder Punktdiagramm ist die ideale Darstellungsform

Korrelation im Streudiagramm veranschaulichen



Der Korrelationskoeffizient

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 * s_y^2}}$$

s_{xy}: Standardabweichung von X und Y

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

s_x^2: Standardabweichung von X

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

s_y^2: Standardabweichung von Y

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

Übungsaufgabe Korrelationskoeffizient

Sie haben folgende Werte gemessen und möchten wissen, wie stark sie miteinander korrelieren:

$x_1 = 1; x_2 = 1; x_3 = 2; x_4 = 2$ und $y_1 = 3; y_2 = 2; y_3 = 1; y_4 = 0$

1. Errechnen Sie den Korrelationskoeffizient.
2. Zeichnen Sie ein Streudiagramm mit den Werten.
3. Interpretieren Sie den Zusammenhang.

Achtung: In der Realität würde man nicht den Zusammenhang von nur 4 Variablen errechnen.

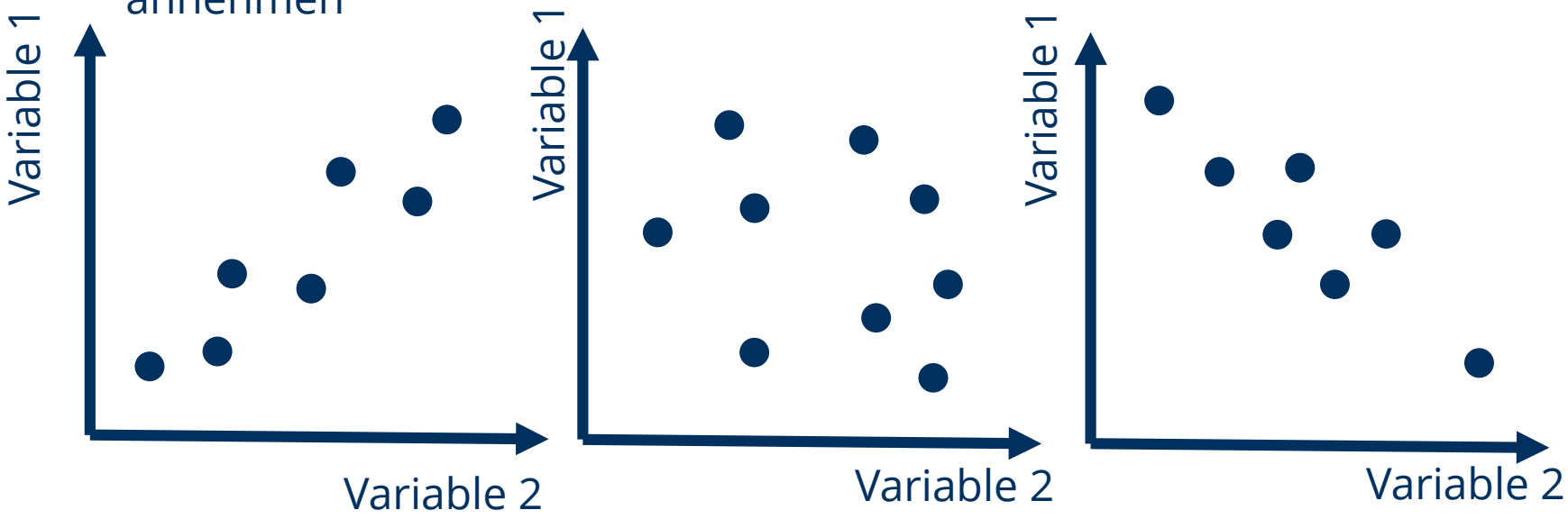
Übung Korrelationskoeffizient: Berechnung

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	3					
1	2					
2	1					
2	0					
$\bar{x} =$ 2,5	$\bar{y} =$ 2,5			$s_x^2 = \sum oben =$	$s_y^2 = \sum oben$	$s_{xy} = \sum oben =$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 * s_y^2}} = \qquad =$$

Korrelationskoeffizient interpretieren I

Der Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen



Positiver Anstieg

Korrelationskoeffizient
z.B. +0,800

Korrelationskoeffizient
z.B. +0,100

Negativer Anstieg

Korrelationskoeffizient
z.B. -0,800

Korrelationskoeffizient interpretieren II

•Nimmt Werte von -1 bis 1 an.

•Bedeutung des Vorzeichen:

Plus (+)	= positiver Zusammenhang
Minus (-)	= negativer Zusammenhang

•Bedeutung des Koeffizienten:

$r < 0{,}3$	kein Zusammenhang
$0{,}3 \leq r < 0{,}5$	schwacher Zusammenhang
$0{,}5 \leq r < 0{,}7$	moderater Zusammenhang
$0{,}7 < r < 0{,}9$	starker Zusammenhang
$0{,}9 < r < 1$	sehr starker Zusammenhang
$r = 1$	perfekter Zusammenhang (unrealistisch in der Wirklichkeit)

Siehe auch Kronthaler, F. (2016): Statistik angewandt: Datenanalyse ist (k)eine Kunst Excel Edition Springer Spektrum.

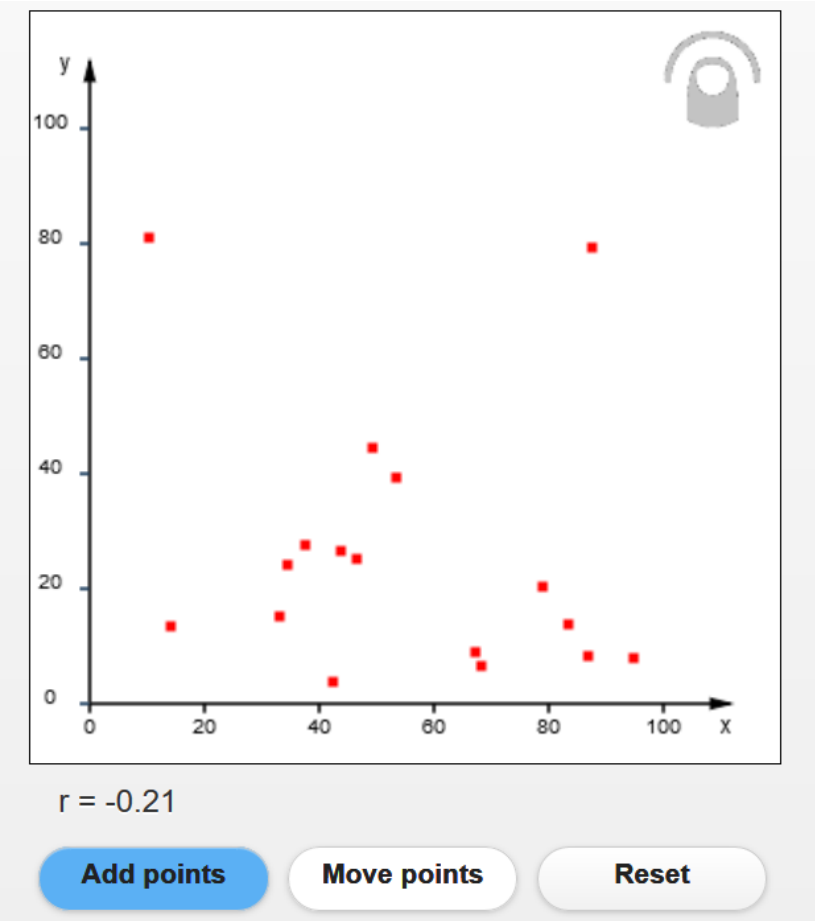
Achtung: Interpretieren Sie den Korrelationskoeffizienten nur wenn der p-Wert unter dem Signifikanzniveau von 0,05 liegt

Übung Korrelationskoeffizient interpretieren

Aufgabe: Öffnen Sie den Link.

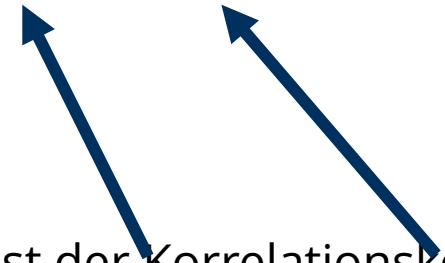
Erstellen Sie sich ein eigenes Streudiagramm und sehen Sie, welchen Korrelationskoeffizient (r) errechnet wird:

https://www.mittag-statistik.de/app/correlation.html



Korrelation: Formel in Excel I

=KORREL(Matrix1;Matrix2)



Variable 1 Variable 1

Das Ergebnis ist der Korrelationskoeffizient (r). Er kann -1 bis +1 reichen.

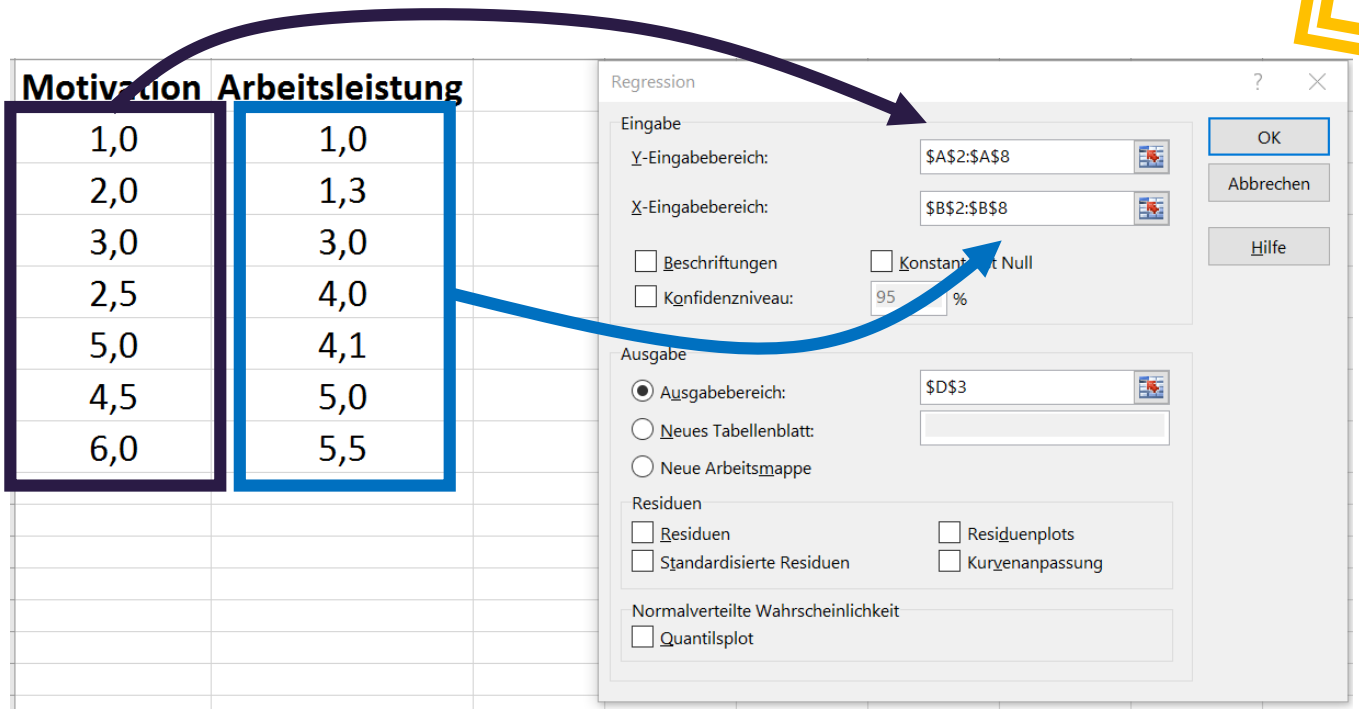
Ein p-Wert wird uns nicht dazu ausgegeben. Weshalb wir in Excel den Umweg über das Datenanalysemodul Regression gehen müssen.

Korrelation: Umsetzung über Datenanalysefunktion

Weil die einfache Formel =KORREL() keinen p-Wert herausgibt müssen wir einen Umweg gehen.

→ Daten → Datenanalyse → Regression

Das Excel-Add-In Datenanalyse muss aktiviert sein



Korrelation: Output interpretieren

Korrelation s-koeffizient

Regressions-Statistik	
Multipler Korrelationskoeffizient	0,892
Bestimmtheitsmaß	0,796
Adjustiertes Bestimmtheitsmaß	0,755
Standardfehler	0,886
Beobachtungen	7

Interpretation: Die Variablen Motivation und Arbeitsleistung korrelieren sehr positiv stark miteinander. Der Korrelationskoeffizient beträgt 0,918, der dazugehörige p-Wert liegt unter dem Signifikanzniveau von 0,05, dies bedeutet, die Nullhypothese – das keine Zusammenhang vorliegt – wird zugunsten der Alternativhypothese zurückgewiesen. Je höher die Motivation ist, desto höher ist auch die Arbeitsleistung.

ANOVA

	Freiheitsgrade (df)	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
Regression	1	15,288	15,288	19,471	0,007
Residue	5	3,926	0,785		
Gesamt	6	19,214			

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	0,295	0,785	0,376	0,723	-1,724	2,313
X Variable 1	0,918	0,208	4,413	0,007	0,383	1,453

p-Wert

Hinweise und Voraussetzungen zum Korrelationskoeffizienten

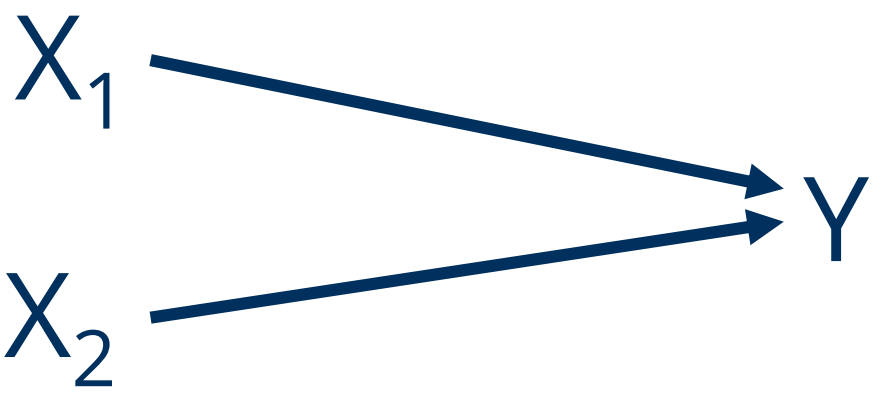
- Trivial: Zu jedem X muss ein Y existieren.
- Diskutierbar: Metrische Variablen werden vorausgesetzt.
- Stichprobegröße: n ≥ 30 *(oder bivariat normalverteilte Variablen: für jedes x sind die y normalverteilt)*
- Lineare Zusammenhänge: Nicht lineare Zusammenhänge werden nicht entdeckt.
- Vorsicht Scheinkausalität: Korrelation bedeutet nicht Kausalität. (https://scheinkorrelation.jimdofree.com/)
- Ausreißer können einen großen Einfluss haben: das testen wir: https://www.mittag-statistik.de/app/correlation.html
- Hier Personsche Korrelationskoeffizient behandelt. Häufige Alternative: Spearmansche Rangkorrelationskoeffizient: Wie lautet die Formel?

BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

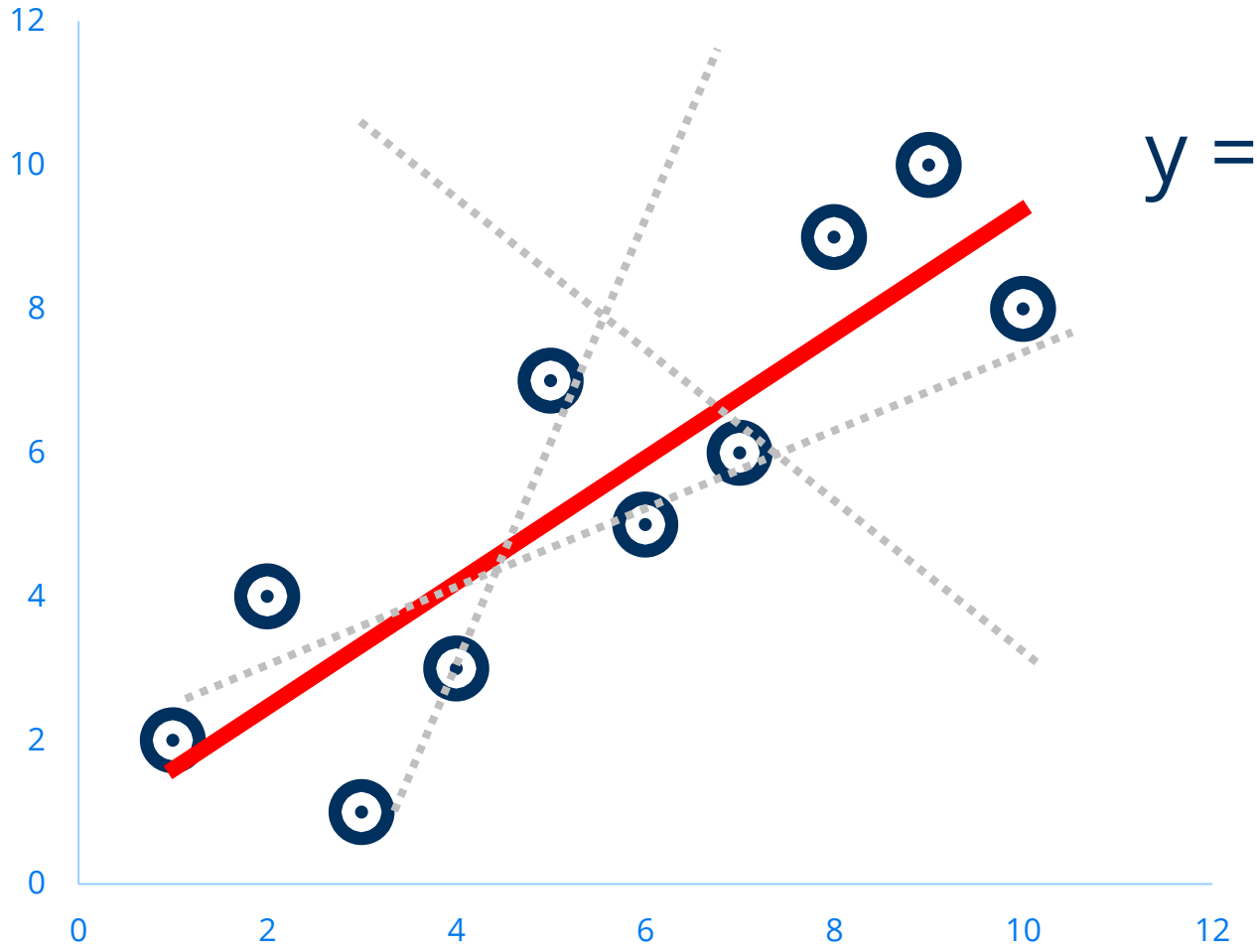
Regression

Lineare Regression Einführung I

Sinn: Eine Variable soll durch zwei oder mehrere andere Variablen erklären. (Dabei wird auch die Bedeutung der einzelnen Variablen analysiert).

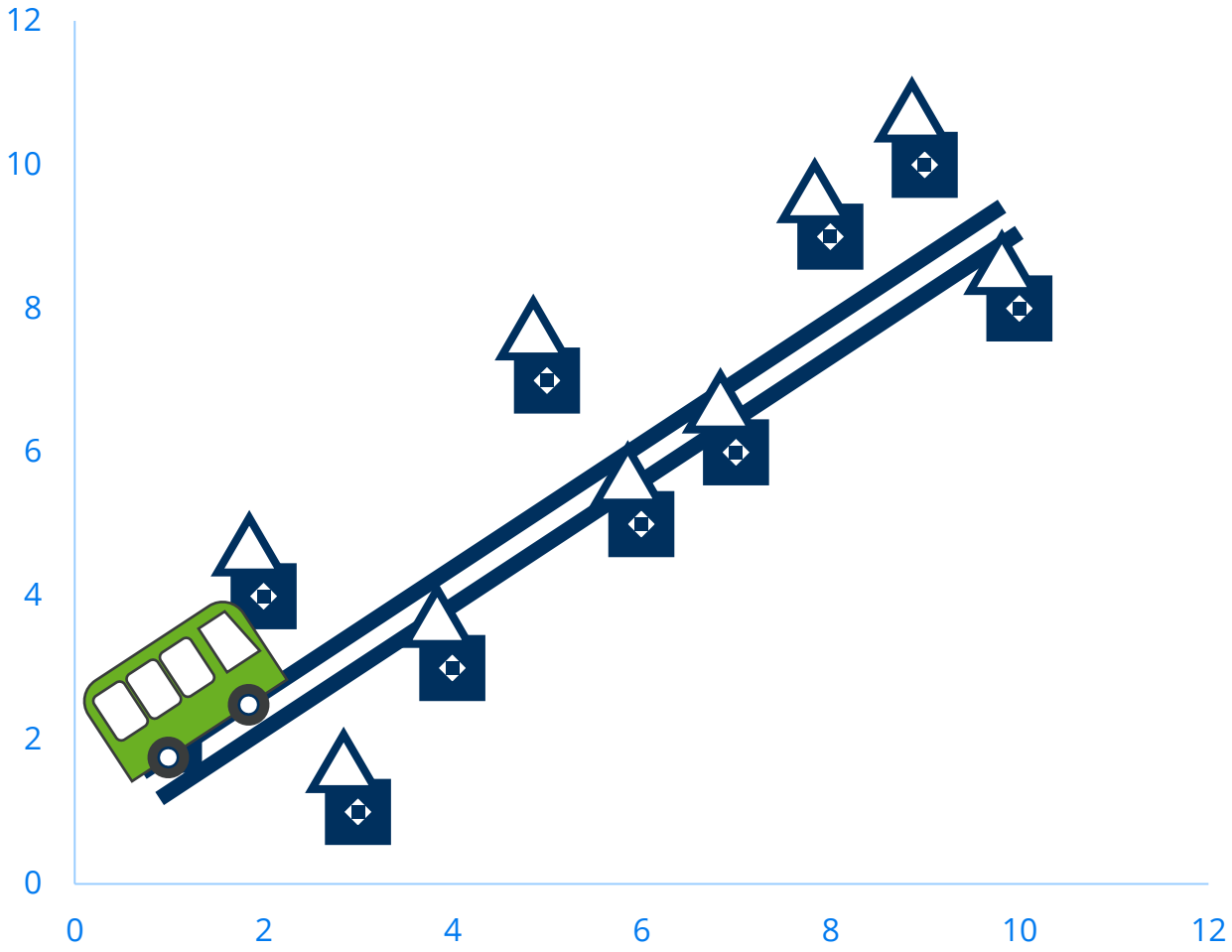


Regressionsgerade

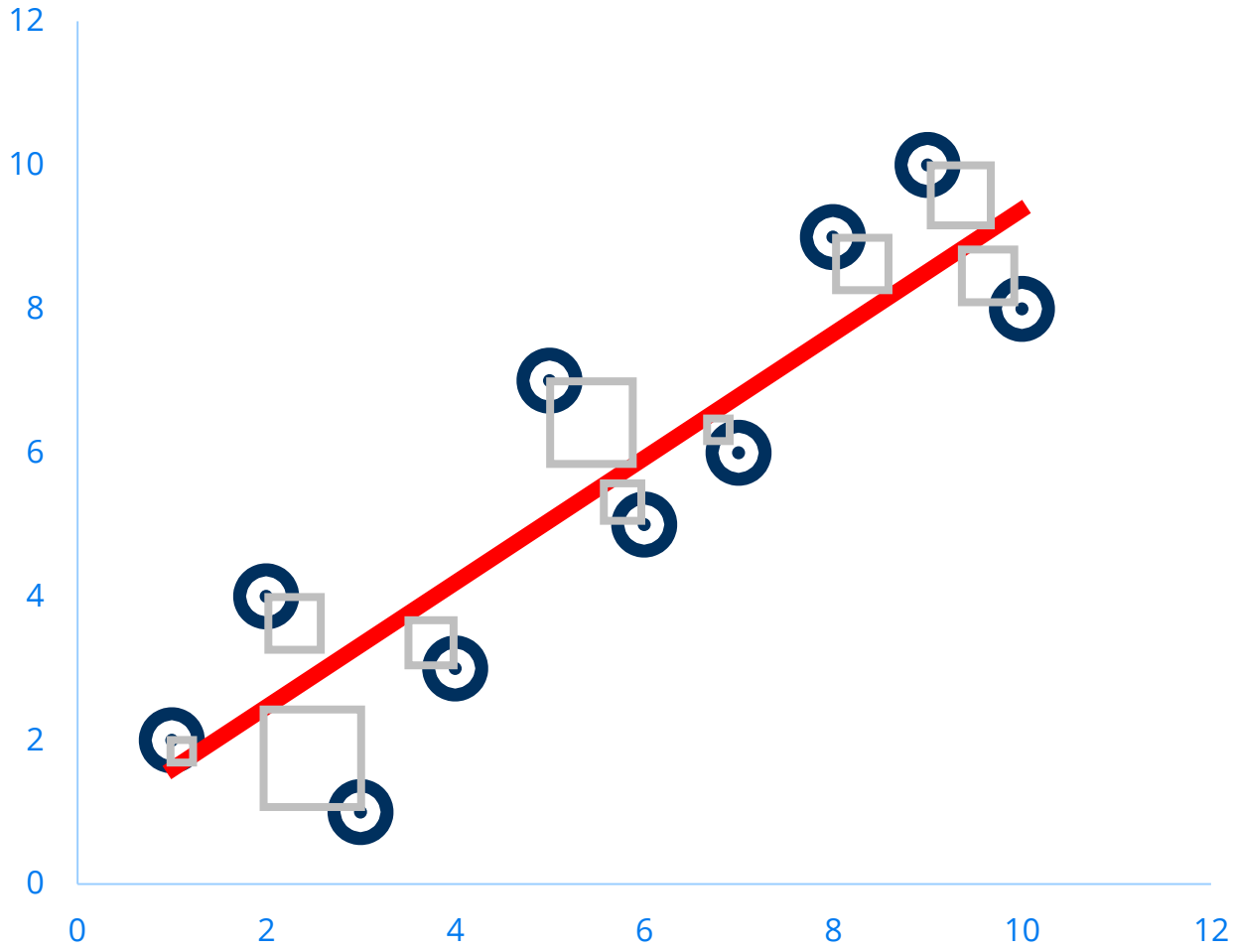


$$y = a + bx$$

Regressionsgerade



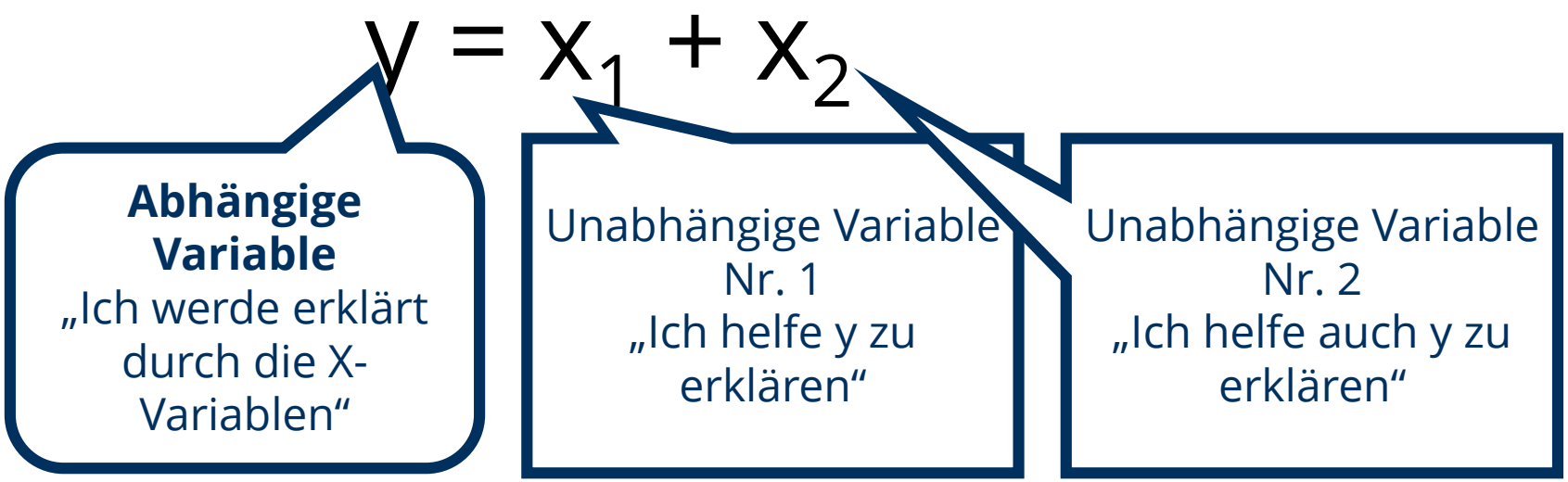
Methode der kleinsten (Fehler) Quadrate



Interaktive online Version: https://seeing-theory.brown.edu/regression-analysis/index.html#section1

Lineare Regression: Einführung II

Allgemein mathematisch ausgedrückt:



Beispiel: Einkommen(y) wird bestimmt(=) durch soziale Schicht(x_1) und(+) Bildungsstand(x_2)

Lineare Regression: Einführung III

Exakter sieht die Formel so aus

$y = Konstante + x_1 * \beta_1 + x_2 * \beta_2 + Fehler$

Ich bin nichts weiter als der Schnittpunkt mit der Y-Achse. Oder die Ausprägung von y wenn x = 0. In den Sozialwissenschaftlichen Analysen ignoriert man mich.

β („Beta“)

Ich gebe an wie stark x Einfluss auf y hat. Negative Zahlen bedeuten negative Zusammenhänge. Zahlen nahe Null sind ein geringer Einfluss.

β_2

Jedes x hat auch ein Beta. Weil jedes x ein unterschiedlich starken Einfluss haben kann auf y.

Ich bin der Teil der nicht erklärt wird durch die Konstante oder die Betas und x‘s. Zur Einführung in die Statistik kann man mich ignorieren.

Regressionsgerade errechnen mit Tabellengleichung

$$y = a + bx$$

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

$$b = \frac{s_{xy}}{s_{xx}}$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
x_1	y_1				
x_2	y_2				
x_3	y_3				
$\bar{x}$	$\bar{y}$			$s_{xx} = \sum$	$s_{xy} = \sum$

Die Summen der quadratischen Abweichungen nun in die Gleichung einsetzen.

Lineare Regression: Wichtige Werte

Bestimmtheitsmaß (R^2):

Nimmt Werte von 0 – 1 an. Die Bedeutung:

0,0 : keine Einfluss der x‘s auf y.

0,5 : mittlerer Einfluss der x‘s auf y.

0,1 : extrem starker Einfluss x‘s auf y.

Beta:

Nimmt Werte von -1 bis +1 an. Die Bedeutung:

-1 : starker negativer Einfluss von x auf y

+0 : kein Einfluss von x auf y

+1 : starker positiver Einfluss von x auf y

Erstellen Sie Ihr eigenes Streudiagramm und sehen Sie welche Werte die Gleichungen annimmt: https://www.mittag-statistik.de/app/regression.html

Übung Regressionsgleichung

x_i	y_i
1	1
2	2
3	2
4	3

Es sind die Werte aus der Tabelle links gegeben. Wie lautet die Regressionsgleichung (a und b)?

$y = a + bx$

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

$b = \frac{s_{xy}}{s_{xx}}$

Nutzen Sie zur Berechnung die Tabellengleichung.

Wie stark ist der Zusammenhang?

Welcher Y-Wert kann für ein X = 5 vorhergesagt werden?

Regressionsgerade errechnen: Übung

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	1				
2	2				
3	2				
4	3				
$\bar{x}$= 2,5	$\bar{y}$=2			$s_{xx} = \sum oben$=	$s_{xy} = \sum$ oben =

$$b = \frac{s_{xy}}{s_{xx}} = \frac{3}{5} =$$

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

$$y = a + b * x =$$ (Regressionsgleichung)

Vorhersage für x =5

$$y(5) =$$ =

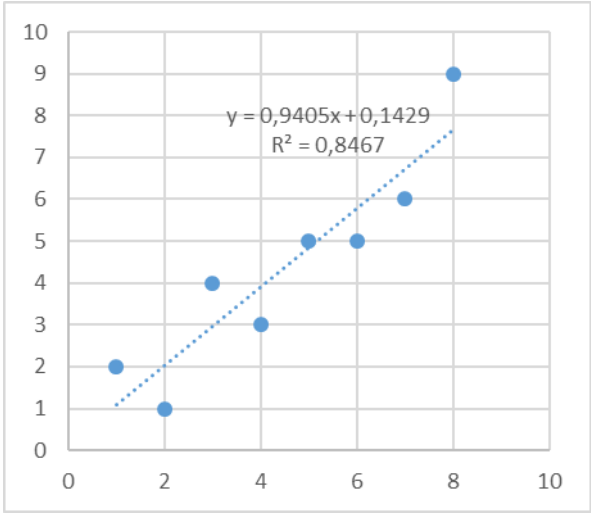
Wenn x = 5 dann y =

Lineare Regression: Darstellung in Excel

Das Vorgehen ist fast identisch mit den Vorgehen, wie wir es bereits für die Korrelation behandelt haben.

Die Darstellung erfolgt wieder mit einem Punktdiagramm. Zusätzlich kann man im Punktdiagramm eine Trendlinie darstellen über → Diagrammtools → Entwurf → Diagrammelement hinzufügen → Trendlinie

- Die Trendlinie kann nach Rechtsklick auf Trendlinie formatiert werden → Trendlinie formatieren → Formel in Diagramm anzeigen und Bestimmtheitsmaß im Diagramm darstellen
- Verwendet wird die Methode der kleinsten Quadrate



Lineare Regression: Alternative Berechnung in Excel

Formel	Entspricht dem Wert aus dem Datenanalysemodell	Bedeutet
=Achsenabschnitt(Y_Werte; X_Werte);	„Schnittpunkt“	Schnittpunkt der Regressionsgeraden mit Y-Achse
=Steigung(Y_Werte;X_Werte);	„Koeffizient“ des X-Wertes	Anstieg der Geraden und auch Höhe des Einflusses von X auf Y.
=Bestimmtheitsmass(Y_Werte;X_Werte)	(auch „Bestimmtheitsmass“)	Wie viel von Y wird durch X erklärt.

Lineare Regression: Berechnung in Excel mit 2 unabhängigen Variablen

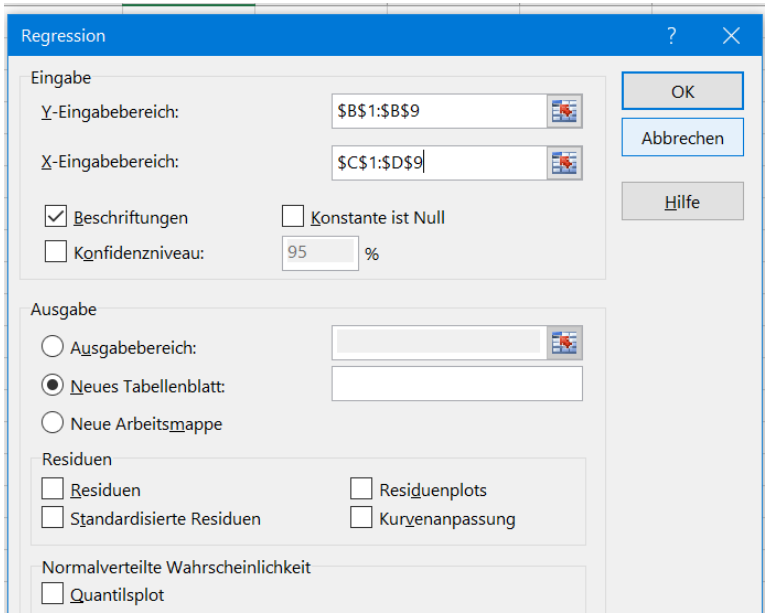
→ Daten → Datenanalyse → Regression

In den Y-Bereich gehört die abhängige Variable

In den X-Bereich gehören die unabhängigen Variablen, dies können auch mehrere sein. (Sie müssen jedoch in Excel in benachbarten Spalten stehen)

Beschriftungen sollten angeklickt sein. Voraussetzung dafür: Sie haben die Spaltenüberschrift auch mitmarkiert

Y	X1	X2
1	2	5
2	1	7
3	4	4
4	3	5
5	5	4
6	5	3
7	6	2
8	9	1



Lineare Regression: Interpretieren I

Der Output zeigt an wie stark y von unseren x's beeinflusst wird

AUSGABE: ZUSAMMENFASSUNG								
Regressions-Statistik								
Multipler Korrelationskoeffizient	0,920							
Bestimmtheitsmaß	0,847							
Adjustiertes Bestimmtheitsmaß	0,785							
Standardfehler	1,135							
Beobachtungen	8							
ANOVA								
	Freiheitsgrade (	*Quad*						
Regression	2							
Residue	5	6,438356164	1,28767123					
Gesamt	7	42						
	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*ntere 95,0%*	*bere 95,0%*
Schnittpunkt	0,658	5,595	0,118	0,911	-13,726	15,041	-13,726	15,041
X1	0,890	0,592	1,503	0,193	-0,633	2,413	-0,633	2,413
X2	-0,014	0,787	-0,017	0,987	-2,036	2,009	-2,036	2,009

Adj. Bestimmtheitsmaß

Nimmt Werte zwischen 0 und +1 an
Sagt wie gut alle x's unser y erklären.
Wie viel Prozent der Varianz (y) kann ich durch mein Modell erklären. (hier 78,5 Prozent der Varianz werden erklärt)
1 = perfekte Erklärung
0 = überhaupt keine Erklärung

Variable 1
Variable 2

Regressionskoeffizient von X1 = 0,890 (je größer, desto größer der Einfluss auf Y, hier großer Einfluss)

Regressionskoeffizient von X2 = -0,014 (X2 hat keine Einfluss auf y)

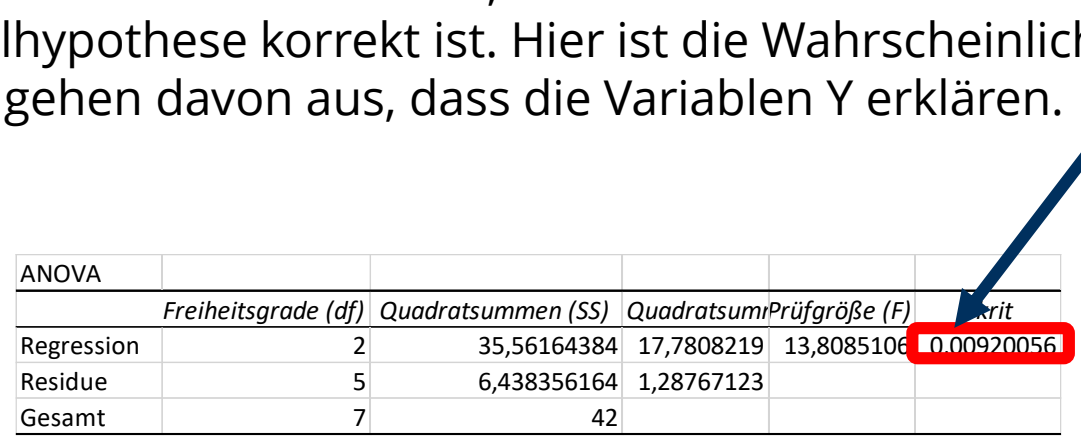
P-Werte der Koeffizienten: Die Irrtumswahrscheinlichkeit. Wenn unter 0,05 = signifikant. Hier sind die p-Werte alle über 0,05 (sogar bei X1 obwohl dort ein starker Koeffizient ist, dies liegt am kleinen n)

Lineare Regression: Interpretieren II

Die ANOVA einer Regression:

Sinn: Erklären die ausgesuchten Variablen das Y. Nullhypothese: Die unabhängigen Variablen erklären Y nicht.

Hier gibt es so etwas ähnliches wie den p-Wert der „F krit“ (kritische Wert). Es ist die Wahrscheinlichkeit, einen solchen F-Wert zu erhalten, wenn die Nullhypothese korrekt ist. Hier ist die Wahrscheinlichkeit sehr klein. Das heißt wir gehen davon aus, dass die Variablen Y erklären.



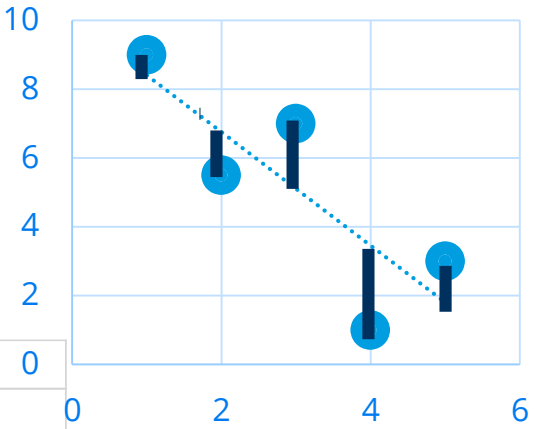
ANOVA					
	Freiheitsgrade (df)	*Quadratsummen (SS)*	*Quadratsumr*	*Prüfgröße (F)*	*krit*
Regression	2	35,56164384	17,7808219	13,8085106	0,00920056
Residue	5	6,438356164	1,28767123		
Gesamt	7	42			

Falls die Regressionskoeffizienten sowieso nicht signifikant waren, ist diese Analyse jedoch sowieso überflüssig. Weshalb sie oft auch gar nicht interpretiert wird.

Lineare Regression: Interpretieren III

Für jeden Fall kann anhand der Regressionsgleichung ein theoretisches Y errechnet werden.

Die Abweichungen vom theoretischen Y und dem tatsächlichen sind die Residuen. (Schwarze Striche links)



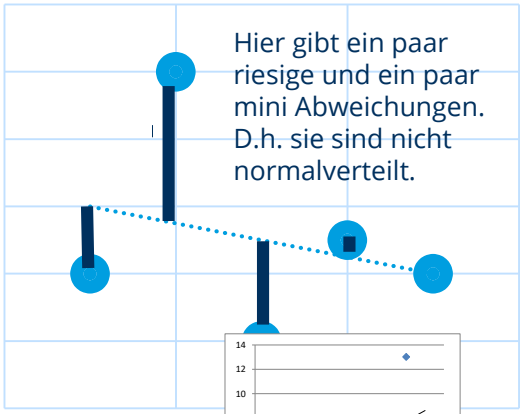
AUSGABE: RESIDUENPLOT			
Beobachtung	*Schätzung für Y*	*Residuen*	*Standardisierte Residuen*
1	2,369863014	-1,369863014	-1,428363228
2	1,452054795	0,547945205	0,571345291
3	4,164383562	-1,164383562	-1,214108743
4	3,260273973	0,739726027	0,771316143
5	5,054794521	-0,054794521	-0,057134529
6	5,068493151	0,931506849	0,971286995
7	5,97260274	1,02739726	1,071272421
8	8,657534247	-0,657534247	-0,685614349

Voraussetzung für Regressionen I

- keine wichtigen unabhängigen Variablen vergessen
- Zusammenhang sollte theoretisch begründet sein
- Multikolinearität: Die unabh. Variablen sollten nicht mit einander stark korrelieren. Bei Multikoll. muss eine Variable aus der Regression entfernt werden – d.h. nicht zwei Variablen verwenden, die eigentlich das gleiche messen („Intelligenz“ und „Klugheit“)
- Die Abweichungen korrelieren nicht mit Variablen – ansonsten erklären diese Variablen Y noch mehr. (Korrelation Residuen mit X‘s)
- Normalverteilte Residuen – nicht dass ein Teil der Y besser erklärt werden und andere große Teile super schlecht (Histogramm der Residuen)

Voraussetzung für Regressionen II

- Aufeinander folgende Abweichungen sollten nicht größer werden – bei Zeitreihen ist das immer der Fall (Autokorrelation; Korrelation Residuen mit Y).
- Die Varianz der Abweichung ist über alle Wert von X gleich (Homoskedastizität; Korrelation Residuen und Y)



Homoskedastizität: auf einem Teil der Gerade weichen die Werte stärker ab als auf dem anderen.

Nützliches Wissen zur Regression

Multiple Regression: nicht 100 Variablen verwenden! Sparsame aber auch erschöpfende Anzahl an Variablen wählen.

Nichtlineare Zusammenhänge: Am besten theoretisch begründet.

Regressionsmodell verändern: nicht relevanten Variablen ausschließen, dies führt jedoch zu neuen Werten (R^2; und Koffezienten). Es sollten nur sinnvolle X's in eine Regression einbezogen werden, für die eine Hypothese vorliegt.

Dummy Variablen: für nominale Daten z.B.

für Geschlecht, wird einfache eine 0 oder 1 verwendet wenn Geschlecht (weiblich) vorhanden ist (1) oder nicht vorhanden ist (0).

Bei einer Variable die drei Kategorien enthält (rot, grün, blau) werden nur zwei Variablen erstellt. Z.B. ROT (0 = rot nicht vorhanden) (1 = rot vorhanden); GRÜN (0 = grün vorhanden) (1 = grün nicht vorhanden). Blau ergibt sich dann logisch wenn ROT = 0 und GRÜN = 0.

Logistische Regression wenn die abhängige Variable nominal ist.

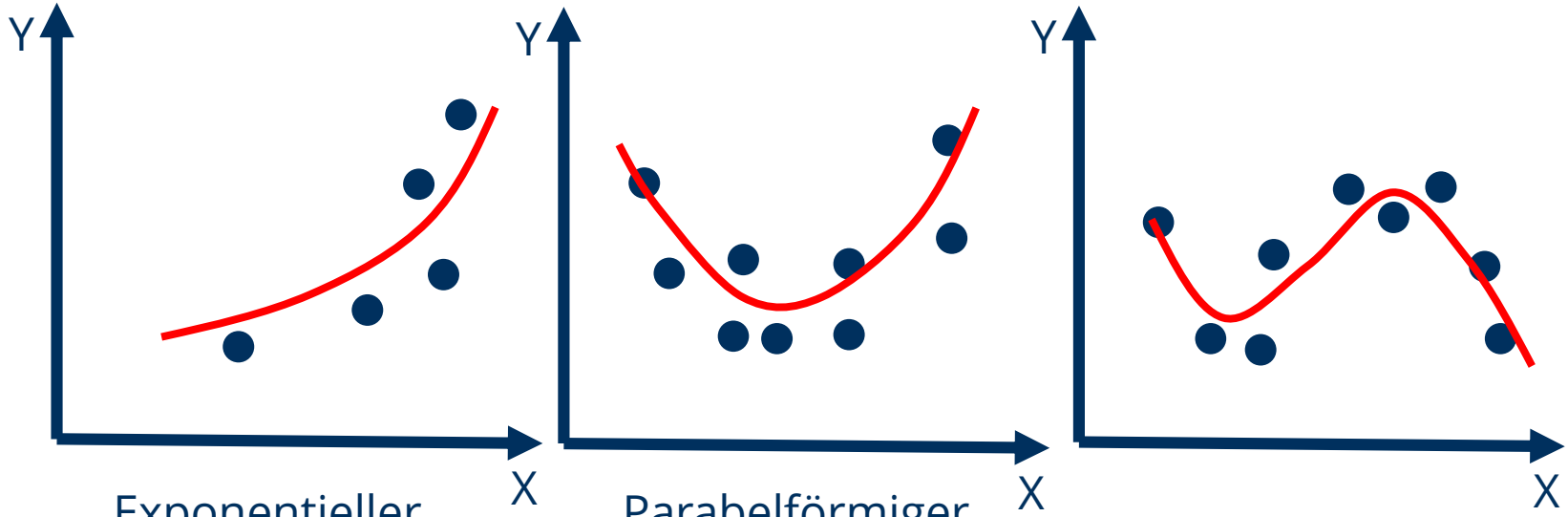
Regression werden auch für Zeitreihenanalysen verwendet).

Lineare Regression erweitern

- Mehr als 1, 2 Variablen sind möglich. Es sollte jedoch nicht 100 Variablen sein! Sparsame aber auch erschöpfende Anzahl an Variablen wählen.
- Ein Regressionsmodell kann abgeändert werden, indem man die nicht relevanten Variablen ausschließt dies führt jedoch zu neuen Werten (R^2; und Koffezienten). Es sollten nur sinnvolle X‘s in eine Regression einbezogen werden, für die eine Hypothese vorliegt.
- Dummy Variablen: für nominale Daten z.B.
- für Geschlecht, wird einfache eine 0 oder 1 verwendet wenn Geschlecht (weiblich) vorhanden ist (1) oder nicht vorhanden ist (0).
- Bei einer Variable die drei Kategorien enthält (rot, grün, blau) werden nur zwei Variablen erstellt. Z.B. ROT (0 = rot nicht vorhanden) (1 = rot vorhanden); GRÜN (0 = grün vorhanden) (1 = grün nicht vorhanden). Blau ergibt sich dann logisch wenn ROT = 0 und GRÜN = 0.
- Logistische Regression wenn die abhängige Variable nominal ist.
- Regression werden auch für Zeitreihenanalysen verwendet).

Regressionen erweitern: nicht lineare Regression

Nichtlineare Zusammenhänge: Theoretisch kann ein Zusammenhang auch nicht linear sein. Dann werden die Daten modelliert: statt z.B. x wird x^2 verwendet.



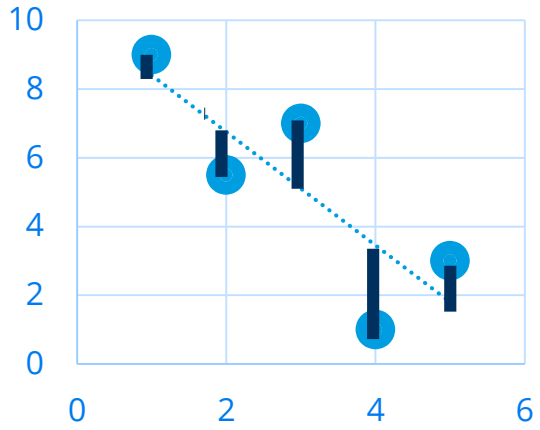
Exponentieller Zusammenhang z.B. $y = a + bx^3$

Parabelförmiger Zusammenhang z.B. $y = a + bx + cx^2$

Polynom z.B. ???

Interpretation der Regressionswerte

- Bei der Interpretation die Codierung der Variablen berücksichtigen: Was bedeutet eine größere Zahl.
- Den Achsenabschnitt nie interpretieren.
- Kausalität beachten: Y wird von den X‘s beeinflusst, nicht umgedreht. Das muss jedoch theoretisch klar sein, der statistische Test kann das nicht beantworten.
- Die Abweichungen vom theoretischen Y und dem tatsächlichen sind die Residuen. (Schwarze Striche links)
- Wir können die Ergebnisse der Regression nutzen, um y-Werte für neue Fälle vorherzusagen (in Sozialwissenschaften macht man das selten).
- Bei der Interpretation die Codierung der Variablen berücksichtigen: Was bedeutet eine größere Zahl.



Zusammenfassung heute