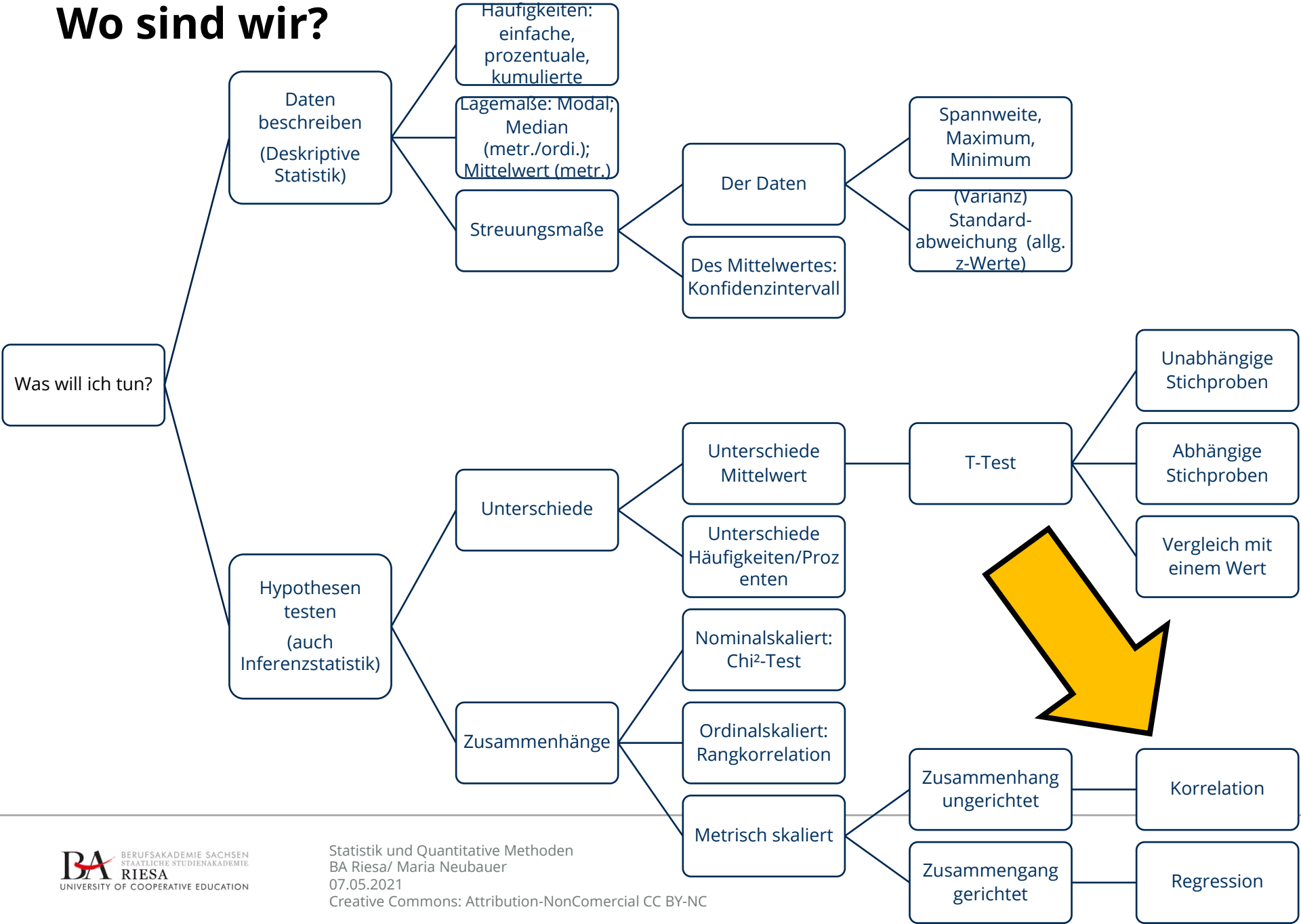
Korrelation

Wo sind wir?

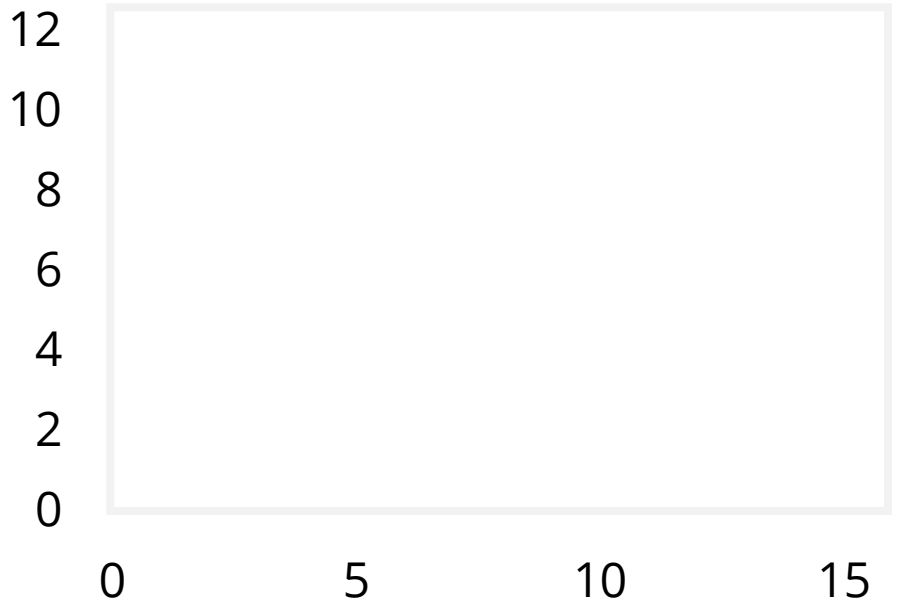


Korrelation: Idee und Streudiagramm

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

Der Korrelationskoeffizient

s_{xy}: gemeinesame Varianz von X und Y

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

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

s_x^2: Varianz von X

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

s_y^2: Varianz von Y

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

Achtung die Varianzen hier unterscheiden sich von der allgemeinen Varianz in der Berechnung.

Beispielaufgabe: 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.

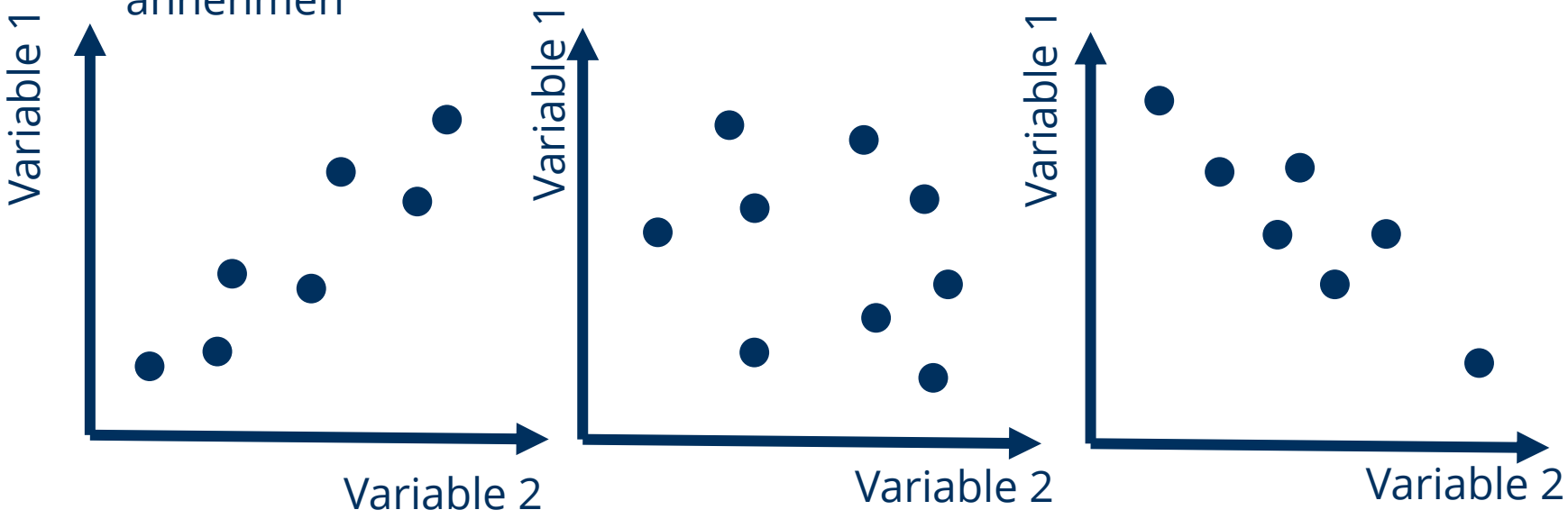
Übung Korrelationskoeffizient: Berechnung

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					
$\bar{x} =$ 1,5	$\bar{y} =$ 1,5			$s_x^2 = \sum oben =$	$s_y^2 = \sum oben =$	$s_{xy} = \sum oben =$

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

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 Korrelationskoeffizienten:

0,0 < 0,1	kein Zusammenhang
0,1 ≤ r < 0,3	geringer Zusammenhang
0,3 ≤ r < 0,5	mittlerer Zusammenhang
0,5 ≤ r <0,9	hoher Zusammenhang
0,9 ≤ r < 1	sehr hoher Zusammenhang
r = 1	Unrealistisch hoher Zusammenhang in der Wirklichkeit

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

Aus Kuckartz et al.: Statistik, Eine verständliche Einführung, 2013, S. 213

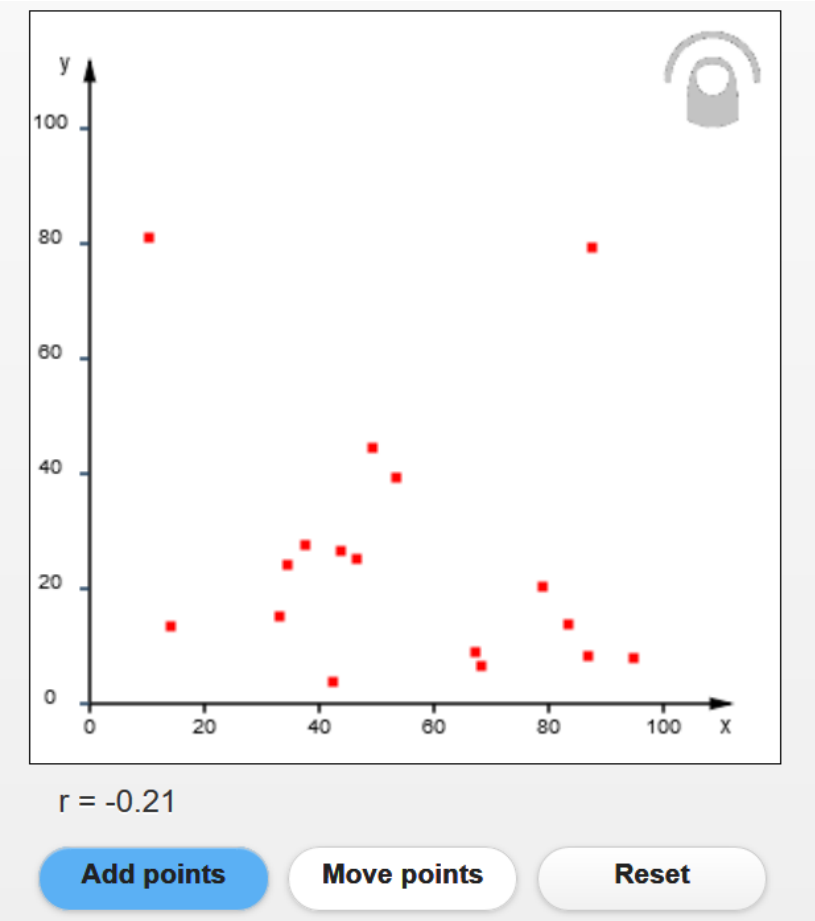
Siehe auch Kronthaler, F. (2016): Statistik angewandt: Datenanalyse ist (k)eine Kunst Excel Edition Springer Spektrum.

Übung Korrelationskoeffizient interpretieren

Aufgabe: Öffnen Sie den Link.

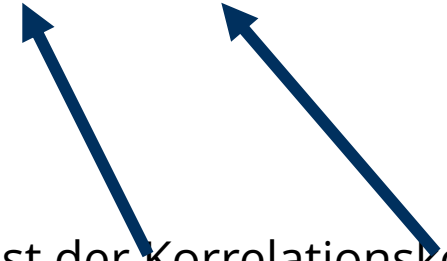
Erstellen Sie sich ein eigenes Streudiagramm und sehen Sie, welchen Korrelationskoeffizient (r) errechnet wird:

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



Korrelation: Formel in Excel I

=KORREL(Matrix1;Matrix2)



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

Variable 1 Variable 1

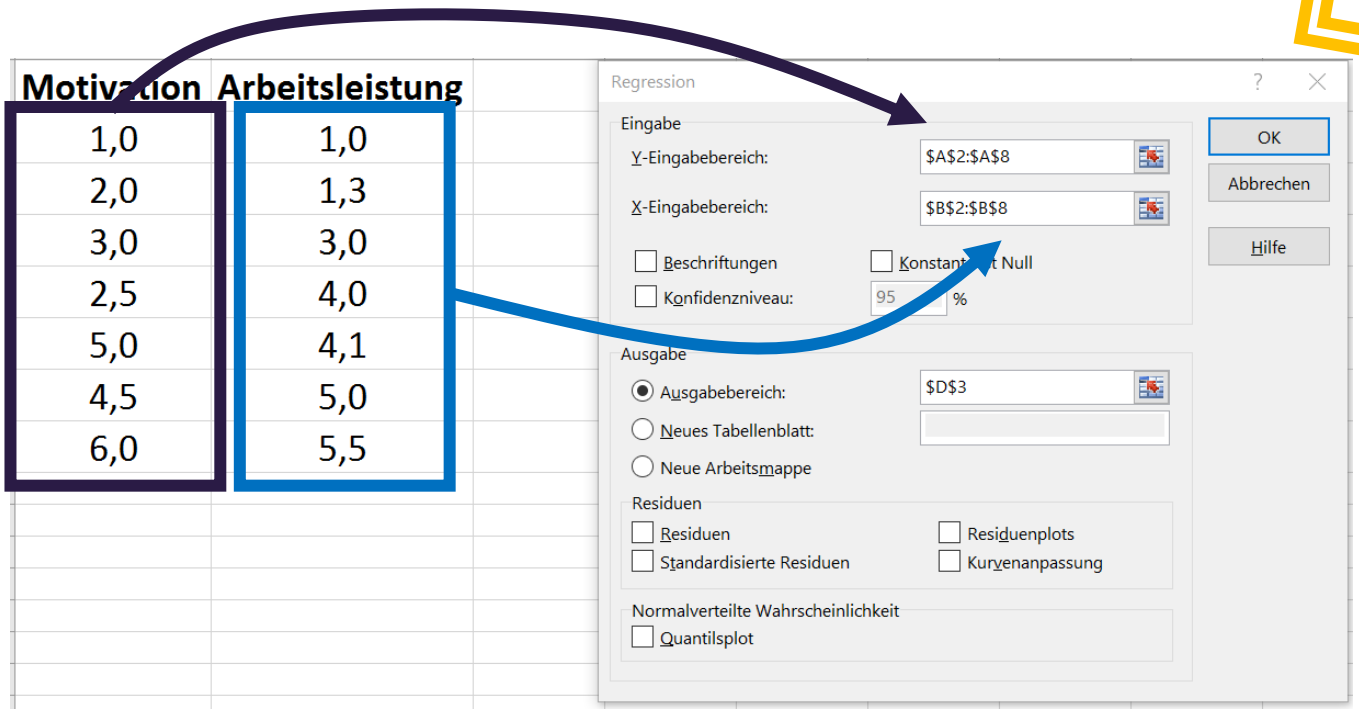
Ein p-Wert wird uns nicht dazu ausgegeben. Weshalb wir in Excel den Umweg über das Datenanalysemodul Regression gehen müssen.

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

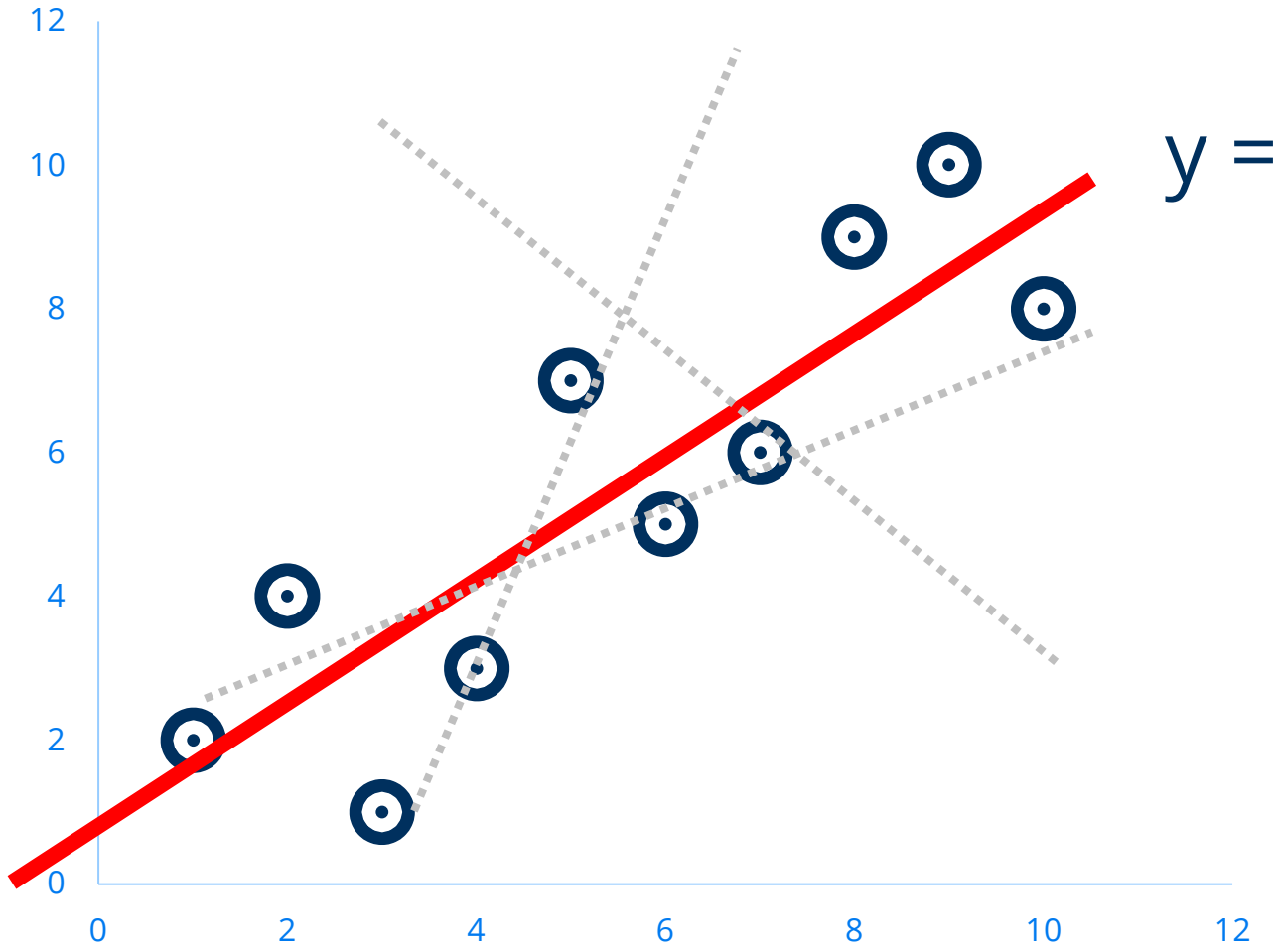
p-Wert

Hinweise und Voraussetzungen zum Korrelationskoeffizienten

- Notwendig: Zu jedem X muss ein Y existieren.
- Diskutierbar: Metrische Variablen werden vorausgesetzt. (Skalen von 1-10 sind auch in Ordnung)
- 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.
- 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?

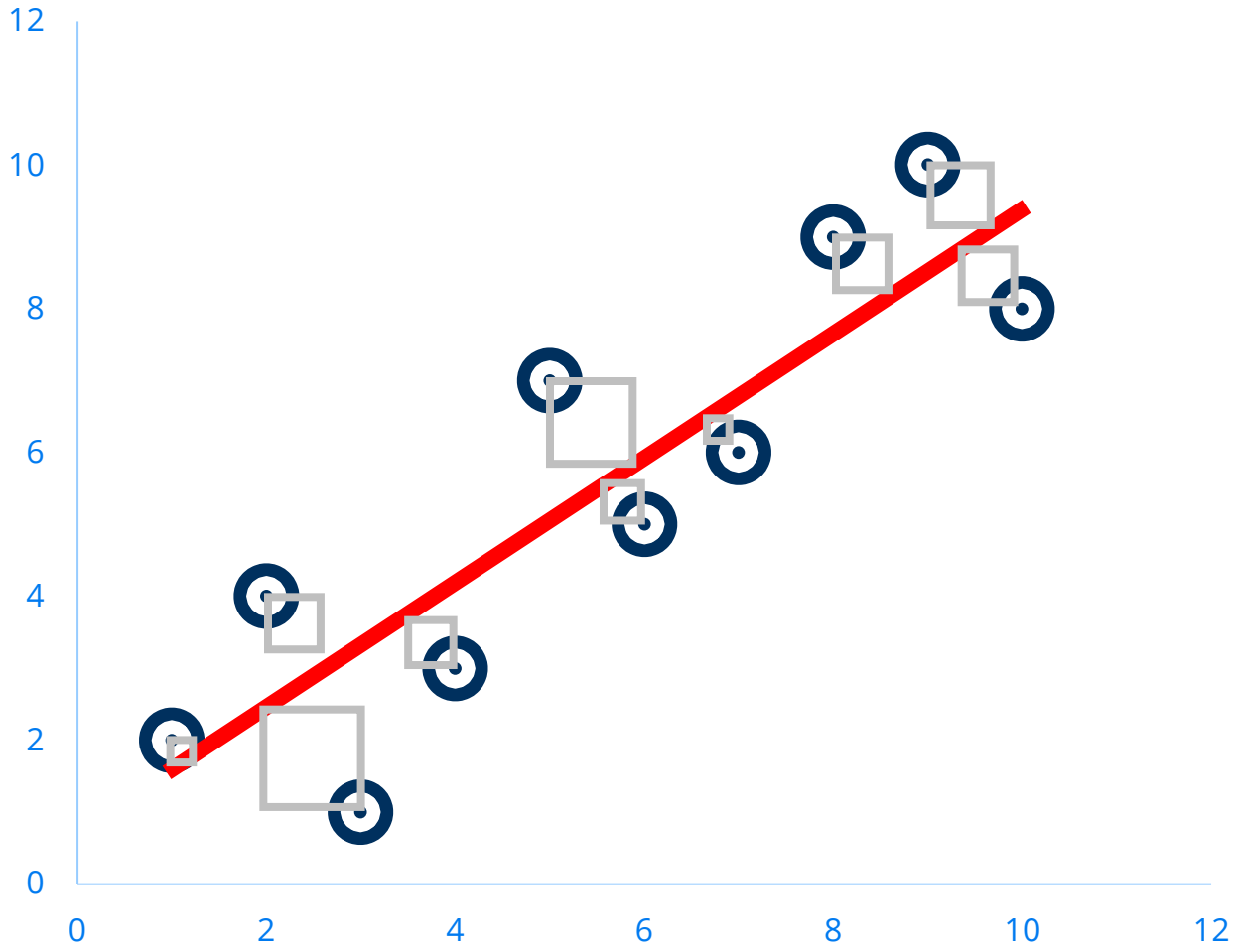
Regression

Regressionsgerade: Wo liegt sie?



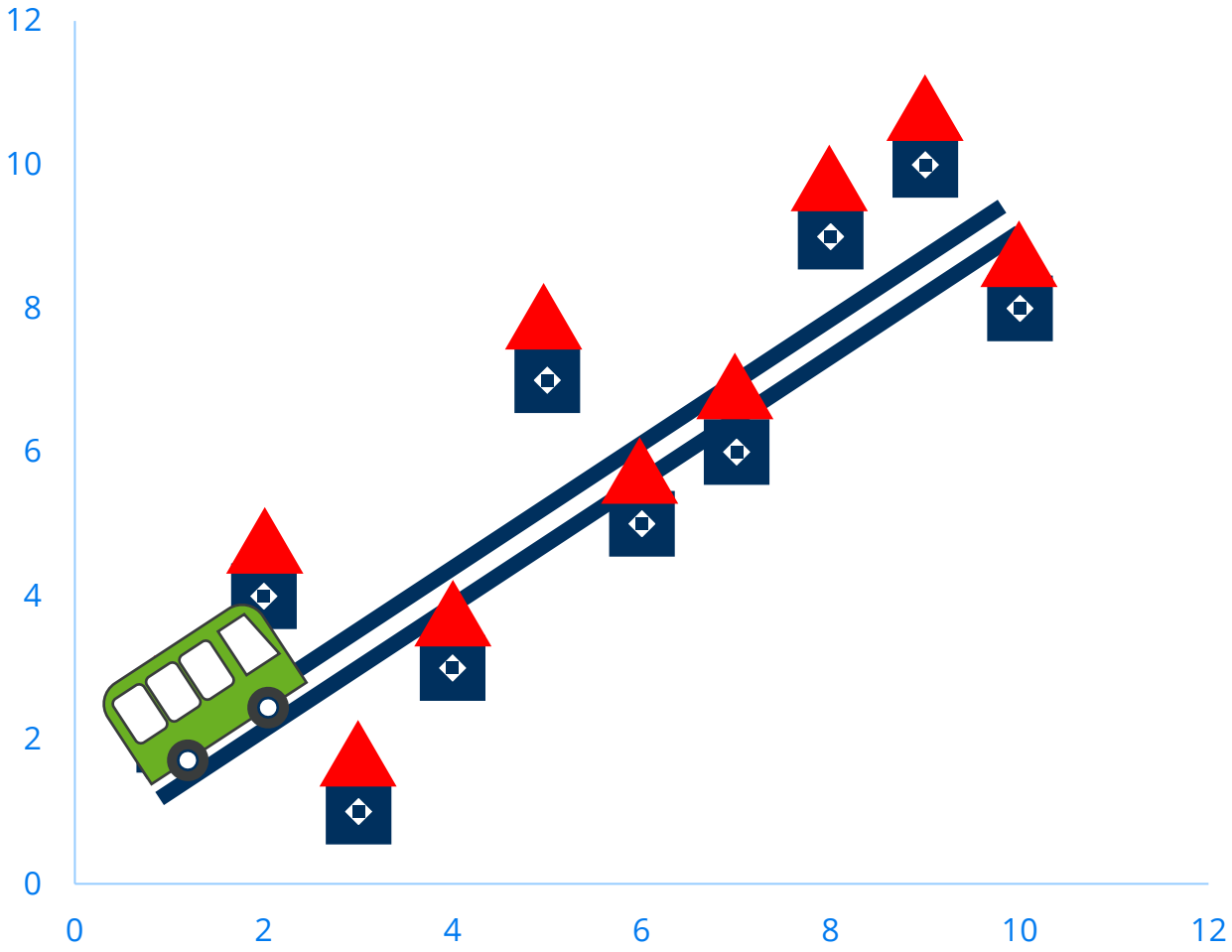
$$y = a + bx$$

Hintergrund für die Regressionsrechnung: Methode der kleinsten (Fehler) Quadrate



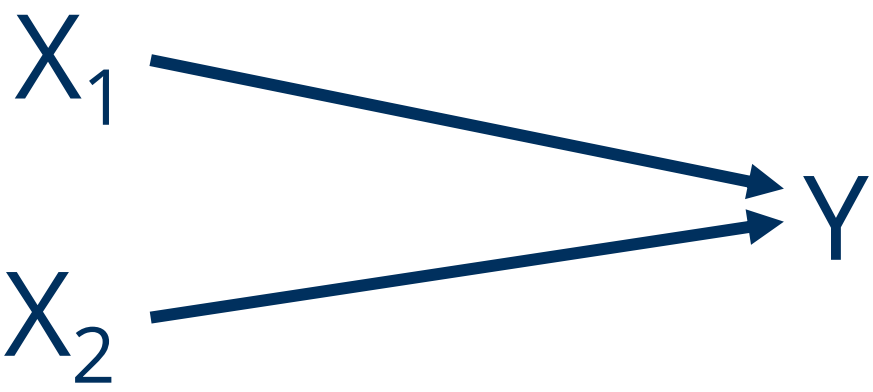
Interaktive online Version: https://seeing-theory.brown.edu/regression-analysis/index.html#section1

Regressionsgerade: Analogie mit einer Busverbindung



Lineare Regression: Ziel

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



Lineare Regression: Abhängige und unabhängige Variablen

Allgemein mathematisch ausgedrückt:

$$y = x_1 + x_2$$

Abhängige Variable
„Ich werde erklärt durch die X-Variablen“

Unabhängige Variable Nr. 1
„Ich helfe y zu erklären“

Unabhängige Variable Nr. 2
„Ich helfe auch y zu erklären“

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

Lineare Regression: Regressionskoeffizienten

Exakter sieht die Formel so aus

$y = \text{Konstante}(a) + x_1 * \beta_1 + x_2 * \beta_2 + \text{Fehler}$

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_1 („Beta“)

Ich gebe an wie stark x_1 Einfluss auf y hat.

β_2
Ich gebe an wie groß der Einfluss von x_2 auf y ist.

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

$$y = a + bx$$

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

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

Formeln noch
Quadratsumm
und
gemeinsam
Abweichung

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: Interpretieren

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.

1 : extrem starker Einfluss x‘s auf y.

(errechnen wir nicht hier)

Interpretation der Regressionskoeffizienten:

Analog zu den Korrelationskoeffizienten

(siehe im Abschnitt zur Korrelation)

Interpretation der Regressionswerte

- Bei der Interpretation die Codierung der Variablen berücksichtigen: Was bedeutet eine größere Zahl.
- Den Achsenabschnitt einer Regressionsgeraden nicht 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.
- Wir können die Ergebnisse der Regression nutzen, um Y-Werte für neue Fälle vorherzusagen (in Sozialwissenschaften macht man das selten).

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 =

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

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

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

Vorhersage für x =5

$$y(5) =$$ =

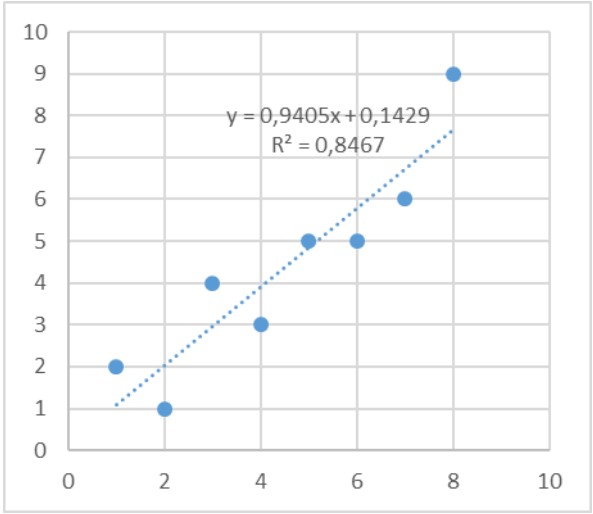
Wenn x = 5 dann y =

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 zwei unabhängigen Variablen

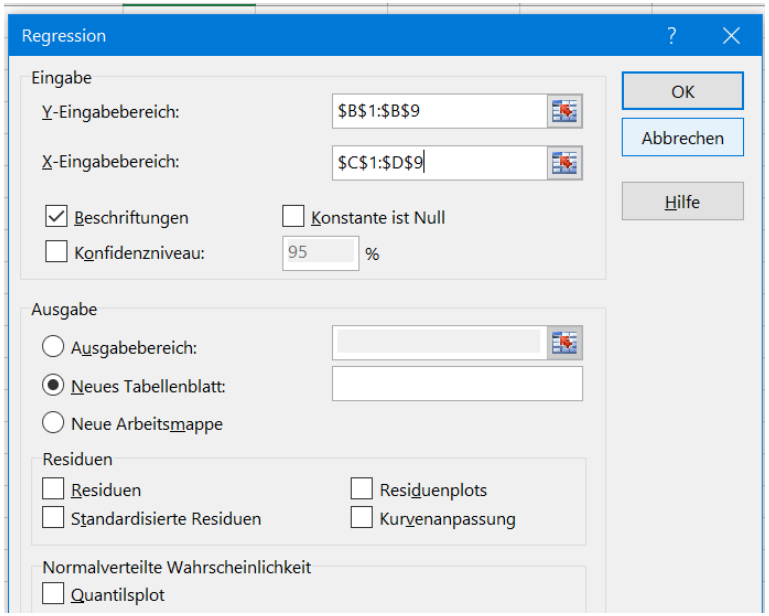
→ Daten → Datenanalyse → Regression

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	8	4
2	9	3
3	6	2
4	7	5
5	5	4
6	4	5
7	3	2
8	2	5
9	1	5



Lineare Regression: Interpretieren I

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

Regressions-Statistik	
Multipler Korrelationskoeffizient	0,970
Bestimmtheitsmaß	0,942
Adjustiertes Bestimmtheitsmaß	0,922
Standardfehler	0,764
Beobachtungen	9

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 92,2 Prozent der Varianz werden erklärt)
1 = perfekte Erklärung
0 = überhaupt keine Erklärung

ANOVA

	Freiheitsgrade (df)	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
Regression	2	56,501	28,251	48,447	0,000
Residue	6	3,499	0,583		
Gesamt	8	60			

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*Untere 95,0%*	*Obere 95,0%*
Schnittpunkt	8,960	1,154	7,765	0,000	6,137	11,783	6,137	11,783
X1	-0,941	0,103	-9,143	0,000	-1,193	-0,689	-1,193	-0,689
X2	0,192	0,222	0,863	0,421	-0,352	0,735	-0,352	0,735

P-Wert 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)

Variable 2 Variable 1

Regressionskoeffizient von X1 = -0,941 (je größer, desto größer der Einfluss auf Y, hier großer Einfluss)

Regressionskoeffizient von X2 = 0,192 (X2 hat keine Einfluss auf y)

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

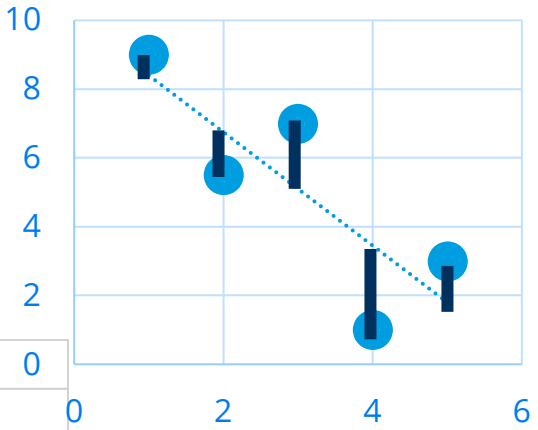
	Freiheitsgrade (df)	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
Regression	2	56,501	28,251	48,447	0,000
Residue	6	3,499	0,583		
Gesamt	8	60			

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), das sind die Fehler.



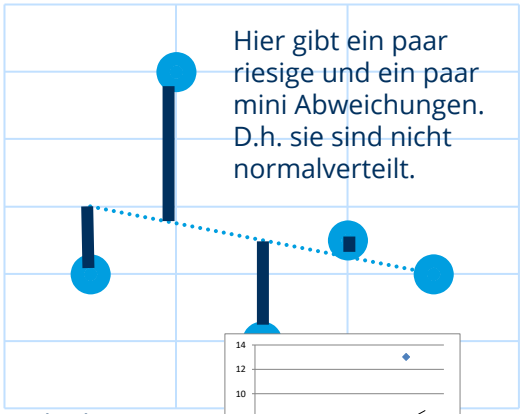
AUSGABE: RESIDUENPLOT			
Beobachtung	*Schätzung für Y*	*Residuen*	*Standardisierte Residuen*
1	2,369863014	-1,369863014	-1,428363228
2	1,452054795	0,547945205	0,571345291
3	4,164383562	-1,164383562	-1,214108743
4	3,260273973	0,739726027	0,771316143
5	5,054794521	-0,054794521	-0,057134529
6	5,068493151	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)

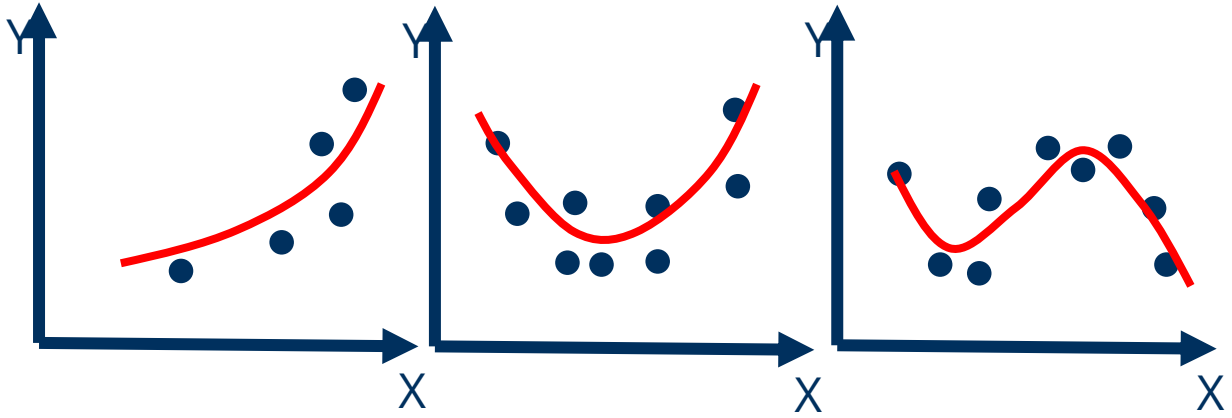


Homoskedastizität: auf einem Teil der Gerade weichen die Werte stärker ab als auf dem anderen.

Weiteres zur Regression

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

Nichtlineare Zusammenhänge: Theoretisch kann ein Zusammenhang auch nicht linear sein. Dann werden die Daten modelliert: statt z.B. x wird x^2 verwendet. Nichtlineare Zusammenhänge: Am besten theoretisch begründet. Kommt in den Sozialwissenschaften selten vor.



Lineare Regression erweitern

- Ein Regressionsmodell kann abgeändert werden, indem man die nicht relevanten Variablen ausschließt dies führt jedoch zu neuen Werten (R^2; und Koeffizienten). 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 (siehe nächsten Folien).

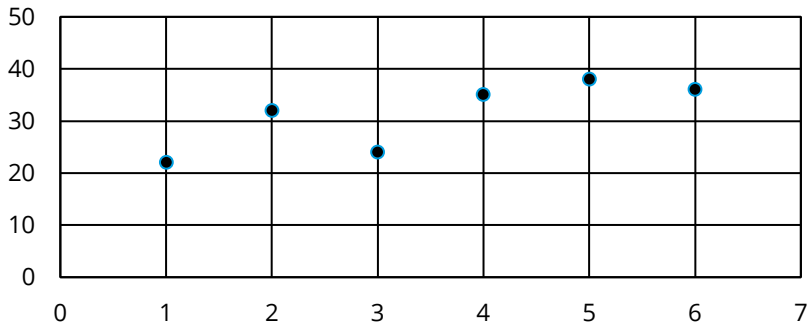
Zeitreihenanalyse

Idee der Zeitreihenanlayse

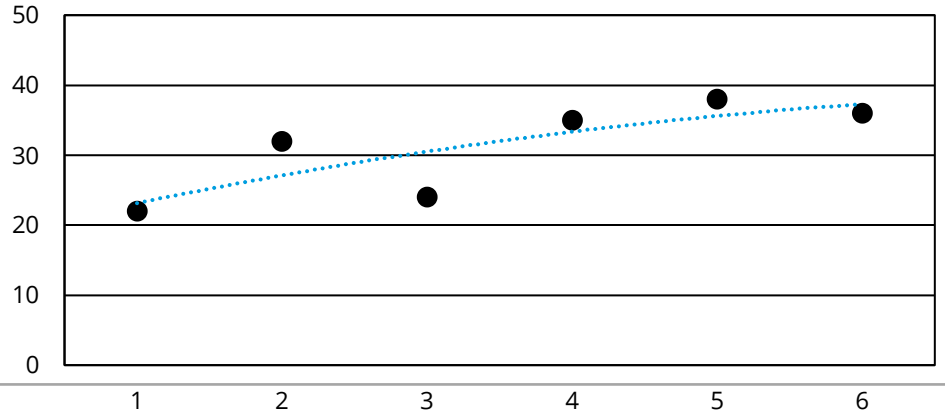
Tabelle mit Werten

Zeitpunkt	Wert
1	22
2	32
3	24
4	35
5	38
6	36

Diagramm der Daten erstellen



Ziel: Trendlinie finden



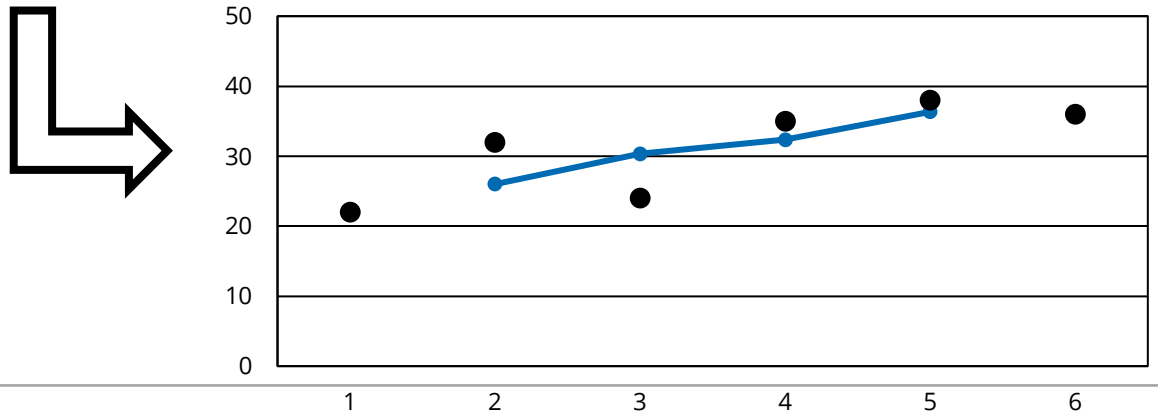
Einfache Trendlinie: Gleitender Durchschnitt

Tabelle mit Werten

Zeitpunkt	Wert	Gleitender Durchschnitt (3. Ordnung)
1	22	
2	32	26,00
3	24	30,33
4	35	32,33
5	38	36,33
6	36	

Mittelwert des vorherigen, aktuellen, und nächsten Zeitpunktes

Gleitender Durchschnitt als Trendlinie



Trendfunktion

$$a = \bar{Y} - b\bar{t} \qquad b = \frac{\sum_{t=1}^{T} tY_t - T\bar{Y}\bar{t}}{\sum_{t=1}^{T} t^2 - T(\bar{t})^2}$$

$\bar{Y} = Mittelwert\ der\ y - Werte = 31{,}16$

$T = Anzahl\ der\ t - Werte, also\ Zeitpunkte = 6$

$$\bar{t} = \frac{Anzahl\ der\ Werte(6) + 1(den\ gesuchten\ Wert)}{2} = 3{,}5$$

$$\sum t\,Y_t = 1 * 22 + 2 * 32 + 3 * 24 + 4 * 35 + 5 * 38 + 6 * 36 = 704$$

$T\bar{Y}\bar{t} = 6 * 31{,}16 * 3{,}5 = 654{,}36$

$$\sum t^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 = 91$$

$T(\bar{t})^2 = 6 * (3{,}5)^2 = 73{,}5$

$$b = \frac{704 - 654{,}36}{91 - 73{,}5} = \frac{49{,}64}{17{,}5} = 2{,}84$$

$a = 31{,}16 - 2{,}84 * 3{,}5 = 21{,}23$

$\hat{m}(\mathrm{t}) = 21{,}23 + 2{,}84t$

Tabelle mit Werten

Zeitpunkt	Wert y_t
1	22
2	32
3	24
4	35
5	38
6	36
7	

$\hat{m}(7) = 21{,}23 + 2{,}84 * 7 = 41{,}11$

Interpretation: Für den Zeitpunkt 7 wird ein Wert von 41,11 vorhergesagt. Das liegt über dem 6. Wert und dem Mittelwert, es ist daher eine Zunahme.