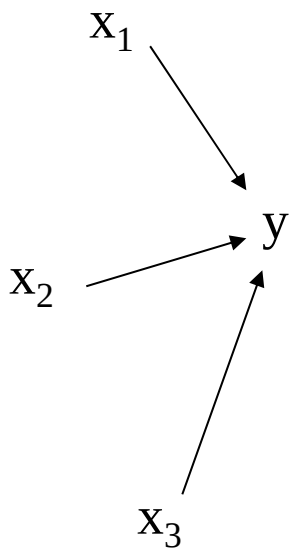
Regression

Ziel der linearen Regression

Bei der linearen Regression wird untersucht, in welcher Weise eine abhängige metrische Variable durch eine oder mehrere unabhängige metrische Variablen durch eine lineare Gleichung beschrieben werden kann.



y...abhängige Variable
(Prognosevariable, Regressand)

$x_1 ... x_j ...$unabhängige Variablen
(Prädiktorvariablen, Regressoren)

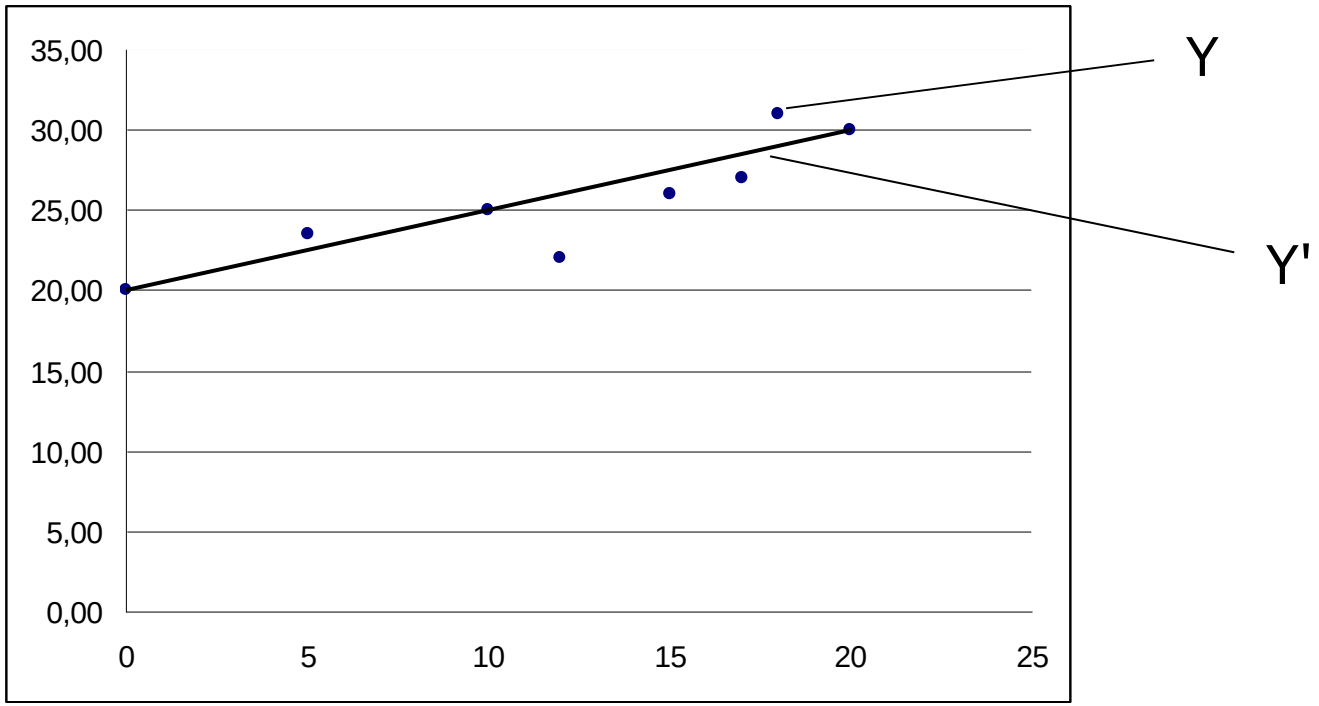
Modellgleichung Regression

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j$$

Y'.....prognostizierte abhängige Variable
$X_{1..i}$....unabhängige Variablen
β_0.....Regressionskonstante
$\beta_{1...i}$....Regressionskoeffizienten der Variablen X_1...X_i

Residuen

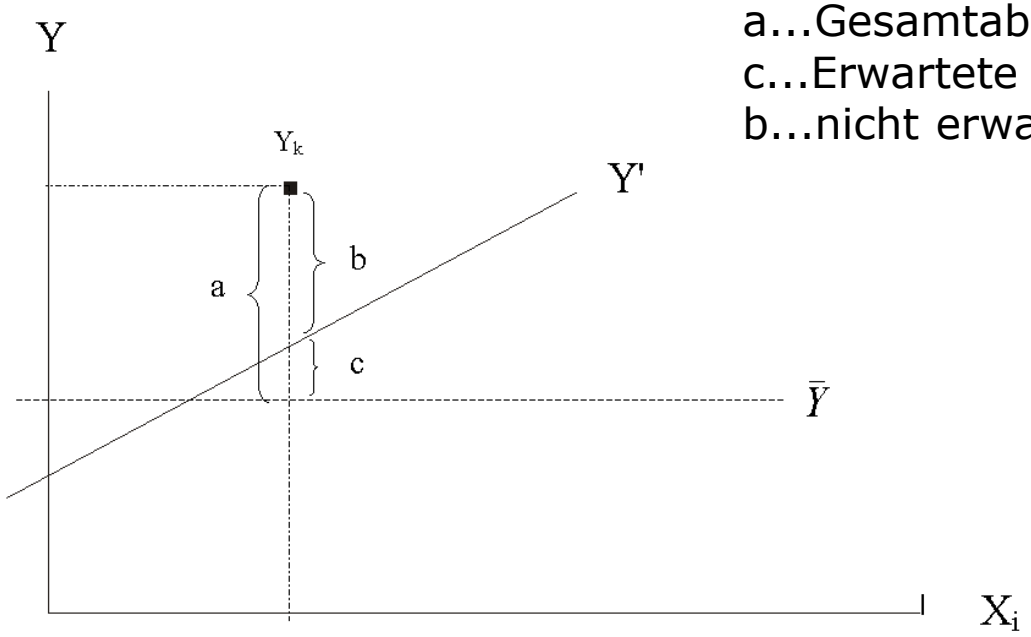
In der Regel verlaufen empirische Wertepaare nicht exakt entlang einer Geraden, sondern weichen mehr oder weniger davon ab.



$Y_k' = \beta_0 + \beta X$ „Prognosewert für Y_k“

$Y_k - Y_k' = e_k$ Residuen

Zielfunktion



a...Gesamtabweichung vom Mittelwert
c...Erwartete Abweichung (aus der Regression)
b...nicht erwartete Abweichung (Residuum)

Zielfunktion der Regression:

$$\sum_{k=1}^{n} e_k^{\,2} = \sum_{k=1}^{n} \left[y_k - \underbrace{(\beta_0 + \beta_1 x_k)}_{} \right]^2 \rightarrow \min!$$

Fehler (b) beobachtet prognostiziert

(Methode der kleinsten Quadrate/Ordinary Least Square, OLS)

Regressionskoeffizienten (Beispiel)

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressionskoeffizient B	Standardfehler	Beta		
1	(Konstante)	-1,964	,673		-2,920	,004
	belast2 Rollenkonflikt	,661	,216	,129	3,056	,002
	belast4 Belastendes SchülerInnenverhalten	,880	,285	,137	3,086	,002
	belast5 Imageprobleme	,495	,189	,111	2,625	,009
	belast7 Physische Belastung	,709	,193	,147	3,679	,000
	belast8 Quantitative Überforderung (Zeitstress)	,915	,249	,161	3,674	,000

a. Abhängige Variable: soma Somatische Beschwerden

Modellgleichung auf Basis der nicht-standardisierten Regressionskoeffizienten:

Soma‘ = -1.964+ B_rolle*0.661 + B_schüler*0.880 + B_image*0.495
+ B_phys*0.709 + B_quanti*0.915

Fortsetzung Beispiel

Skala	Anzahl Items der Skala	Mittelwert
Belastendes Schülerverhalten	9	2.2220
Quantitative Überforderung	5	2.7037
Physische Belastung	3	1.8954
Rollenkonflikte	3	1.7103
Imageprobleme	4	2.8100
Somatische Beschwerden	14	6.3322

Soma‘ = -1.964+ B_rolle*0.661 + B_schüler*0.880 + B_image*0.495
+ B_phys*0.709 + B_quanti*0.915

↓ Mittelwerte für UAV einsetzen

Soma‘ = -1.964+ 1.7103*0.661 + 2.222*0.880 + 2.81*0.495
+ 1.8954*0.709 + 2.7037*0.915=
6.33 = Mittelwert von Somatische Beschwerden

↓ Was passiert wenn z.B. Schülerverhalten um eine Einheit erhöht wird (von 2.222 auf 3.222)?

Soma‘= -1.964+ 1.7103*0.661 + 3.222*0.880 + 2.81*0.495
+ 1.8954*0.709 + 2.7037*0.915 = 7.21

Standardisierte / Nichtstandardisierte Regressionskoeffizienten

Nicht-Standardisierte Regressionskoeffizienten geben an, um wie viele Einheiten sich die abhängige Variable (gemäß der Annahme der Regressionsgleichung) verändert, wenn sich die unabhängige Variable um eine Einheit erhöht.

Nicht-Standardisierte Regressionskoeffizienten sind von den Einheiten der Variablen abhängig und daher untereinander nicht vergleichbar

Daher Standardisierte Regressionskoeffizienten berechnet, welche mit den Standardabweichungen von x und y normiert werden.

$$b_i = \beta_i \frac{s_{x_i}}{s_y}$$

Geben an, um wie viele Standardabweichungen sich y verändert, wenn x um eine Standardabweichung erhöht wird

Eigenschaften der Regressionskoeffizienten

Prognosekoeffizienten:

Gibt an, um welchen Anteil y erhöht wird, wenn x_i um eine Einheit (Standardabweichung) erhöht wird

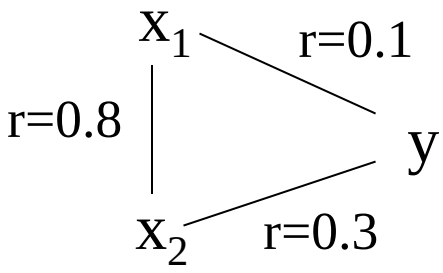
Stärkekoeffizienten:

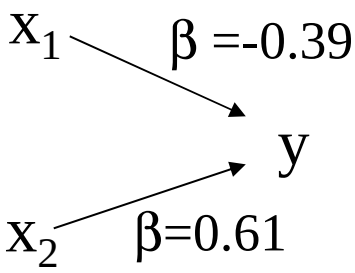
Standardisierte Regressionskoeffizienten sind zwischen 0 und 1 normiert und drücken wie Korrelationskoeffizienten die Stärke eines Zusammenhanges aus.

partielle Koeffizienten:

Gibt den direkten Einfluss von x_i auf y an, wenn die übrigen unabhängigen Variablen konstant gehalten werden.

Direkte – indirekte Zusammenhänge





wäre der direkte Zusammenhang (x_1,y)=0 dann würde x_1 mit y dennoch mit 0.8*0.3=0.24 korrelieren.

Tatsächlich ist die Korrelation jedoch nur 0.1. Daher muss es einen negativen direkten Zusammenhang zwischen x_1 und y geben.

Anmerkung: Im Fall von zwei x-Variablen lassen sich β_1 und β_2 folgenderweise berechnen:

$$\beta_1 = \frac{r_{(y,x_1)} - r_{(y,x_2)} \cdot r_{(x_1,x_2)}}{1 - r^2_{(x_1,x_2)}}, und\ \beta_2 = \frac{r_{(y,x_2)} - r_{(y,x_1)} \cdot r_{(x_1,x_2)}}{1 - r^2_{(x_1,x_2)}}$$

Beispiel

Korrelationen

		soma Somatische Beschwerden	belast2 Rollenkonflikt	belast4 Belastendes Schüler Innenverhalten	belast5 Imageprobleme	belast7 Physische Belastung	belast8 Quantitative Überforderung (Zeitstress)
soma Somatische Beschwerden	Korrelation nach Pearson	1	,317**	,363**	,324**	,327**	,362**
	Signifikanz (2-seitig)		,000	,000	,000	,000	,000
	N	630	619	610	621	623	626
belast2 Rollenkonflikt	Korrelation nach Pearson	,317**	1	,469**	,306**	,273**	,348**
	Signifikanz (2-seitig)	,000		,000	,000	,000	,000
	N	619	650	633	645	646	646
belast4 Belastendes SchülerInnenverhalten	Korrelation nach Pearson	,363**	,469**	1	,358**	,378**	,404**
	Signifikanz (2-seitig)	,000	,000		,000	,000	,000
	N	610	633	641	639	639	635
belast5 Imageprobleme	Korrelation nach Pearson	,324**	,306**	,358**	1	,299**	,492**
	Signifikanz (2-seitig)	,000	,000	,000		,000	,000
	N	621	645	639	653	649	647
belast7 Physische Belastung	Korrelation nach Pearson	,327**	,273**	,378**	,299**	1	,336**
	Signifikanz (2-seitig)	,000	,000	,000	,000		,000
	N	623	646	639	649	655	649
belast8 Quantitative Überforderung (Zeitstress)	Korrelation nach Pearson	,362**	,348**	,404**	,492**	,336**	1
	Signifikanz (2-seitig)	,000	,000	,000	,000	,000	
	N	626	646	635	647	649	656

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

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		Regressionskoeffizient B	Standardfehler	Beta		
1	(Konstante)	-1,964	,673		-2,920	,004
	belast2 Rollenkonflikt	,661	,216	,129	3,056	,002
	belast4 Belastendes SchülerInnenverhalten	,880	,285	,137	3,086	,002
	belast5 Imageprobleme	,495	,189	,111	2,625	,009
	belast7 Physische Belastung	,709	,193	,147	3,679	,000
	belast8 Quantitative Überforderung (Zeitstress)	,915	,249	,161	3,674	,000

a. Abhängige Variable: soma Somatische Beschwerden

Beispiel 2

Korrelationen

		V18 Unterstützungspersonen	V22 Freizeit_personen	V39 gesundheit
V18 Unterstützungspersonen	Korrelation nach Pearson	1	,382**	,287**
	Signifikanz (2-seitig)		,000	,000
	N	245	244	245
V22 Freizeit_personen	Korrelation nach Pearson	,382**	1	,166**
	Signifikanz (2-seitig)	,000		,009
	N	244	244	244
V39 gesundheit	Korrelation nach Pearson	,287**	,166**	1
	Signifikanz (2-seitig)	,000	,009	
	N	245	244	246

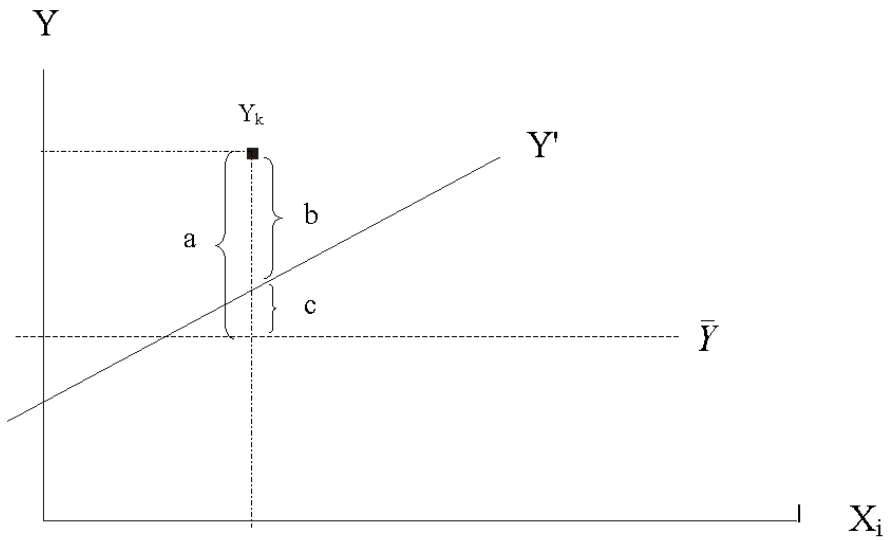
**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	6,217	,238		26,175	,000
	V22 Freizeit_personen	,013	,013	,066	,988	,324
	V18 Unterstützungspersonen	,079	,020	,263	3,944	,000

a. Abhängige Variable: V39 gesundheit

Modellanpassung – R^2



Gesamtstreuung = erklärte Streuung + nicht erklärte Streuung

a = c + b

$$\sum_{j=1}^{n}\left(y_j-\bar{y}\right)^2=\sum_{j=1}^{n}\left(y_j'-\bar{y}\right)^2+\sum_{j=1}^{n}\left(y_j-y_j'\right)^2$$

$$R^2=\frac{erklärteStreuung}{GesamtStreuung}=\frac{\sum_{j=1}^{n}\left(y_j'-\bar{y}\right)^2}{\sum_{j=1}^{n}\left(y_j-\bar{y}\right)^2}$$

R^2

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,479[a]	,229	,223	2,87333

a. Einflußvariablen : (Konstante), belast8 Quantitative Überforderung (Zeitstress), belast7 Physische Belastung, belast2 Rollenkonflikt, belast5 Imageprobleme, belast4 Belastendes SchülerInnenverhalten

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	1449,167	5	289,833	35,106	,000[a]
	Nicht standardisierte Residuen	4871,054	590	8,256		
	Gesamt	6320,221	595			

a. Einflußvariablen : (Konstante), belast8 Quantitative Überforderung (Zeitstress), belast7 Physische Belastung, belast2 Rollenkonflikt, belast5 Imageprobleme, belast4 Belastendes SchülerInnenverhalten

b. Abhängige Variable: soma Somatische Beschwerden

- R^2="Bestimmtheitsmaß" oder "R-Quadrat"

- R^2=1 wenn Gesamtstreuung durch das Modell vollständig aufgeklärt wird. (erklärte Streuung = Gesamtstreuung)

- R^2=0 wenn das Modell nichts an der Varianz von y aufklärt

- R^2*100=Prozentsatz der erklärten Varianz

- $\sqrt{R^2} = r$ =multipler Korrelationskoeffizient = r(y,y‘)

- aus R^2-$R^2_{(i)}$ kann der Anteil erklärter Varianz durch die unabhängige Variable i berechnet werden

Regression - Anwendungsvoraussetzungen

Metrisches Messniveau aller Variablen

Lineare Zusammenhänge zwischen x und y

Normalverteilung und Varianzhomogenität der Residuen

Die Residuen $e_i=y_i-y_i'$ müssen für alle Ausprägungskombinationen der unabhängigen Variablen normalverteilt mit Mittelwert 0 sein.

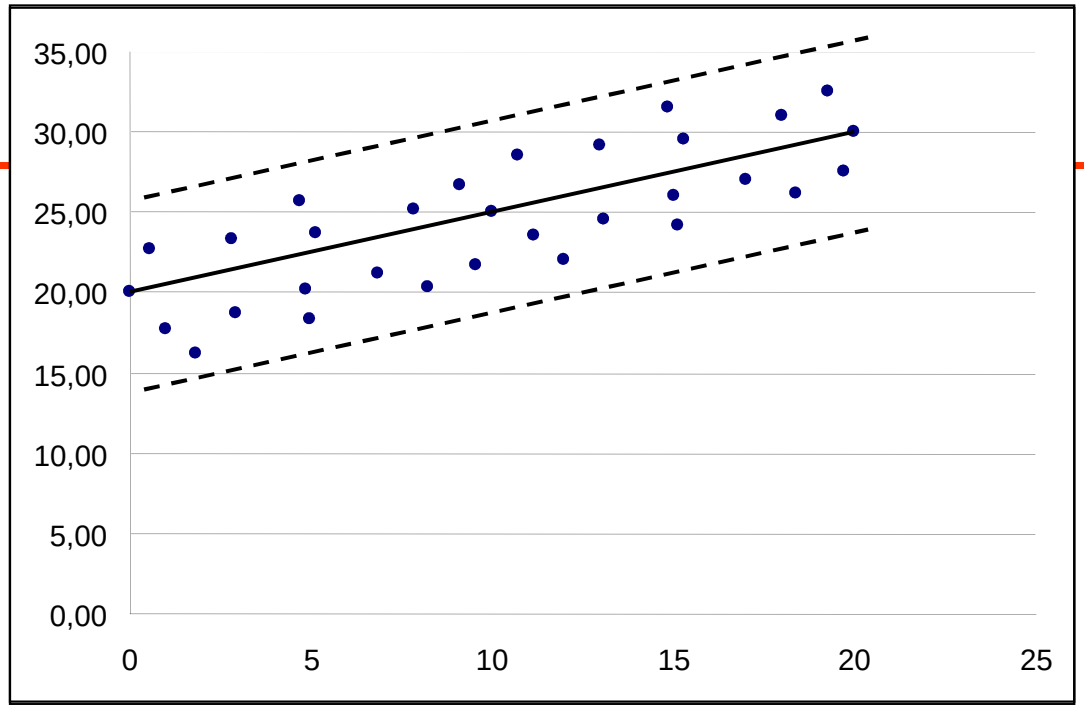
Bei großen Stichproben ist diese Voraussetzung von untergeordneter Bedeutung (z.B. Greene, 2000: 278f).

Keine Multikollinearität

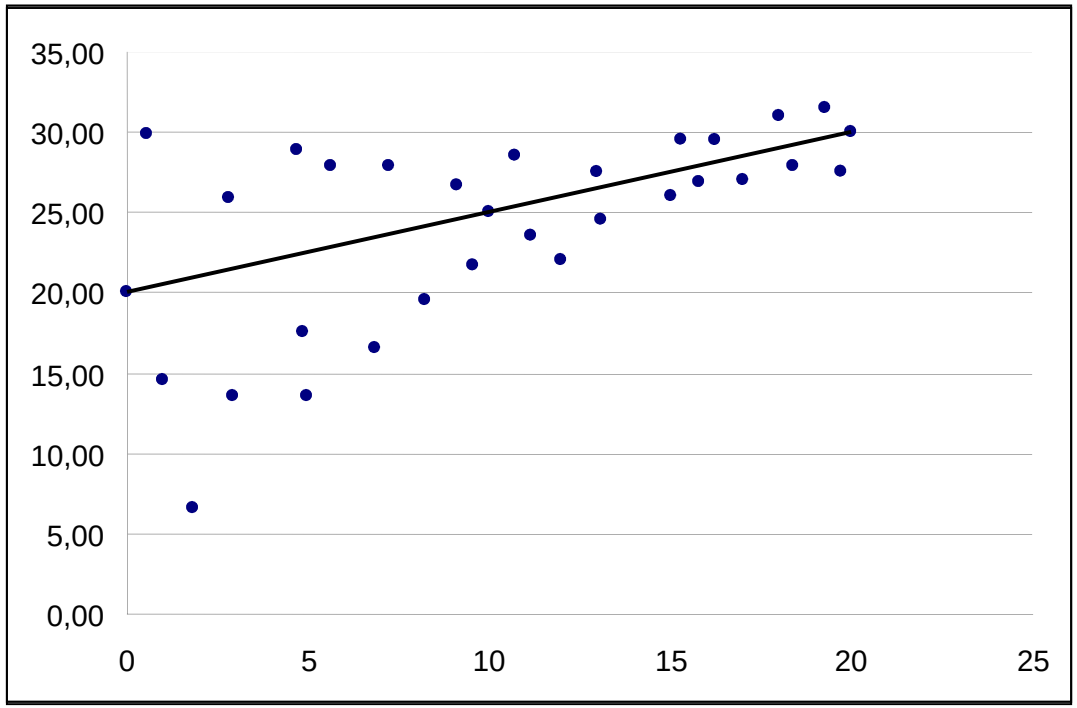
Multikollinearität tritt auf, wenn mindestens eine unabhängige Variable eine Linearkombination einer anderen unabhängigen Variablen darstellt.

„Beinahe-Multikollinearität“ bewirkt ebenfalls eine Verzerrung der Koeffizienten.

Homoskedastizität:



Heteroskedastizität:

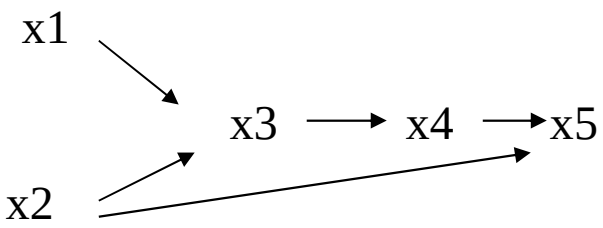


Die Explorative Pfadanalyse

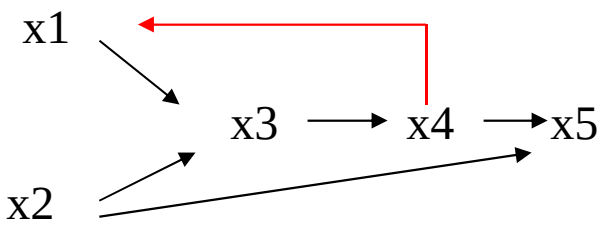
Die Pfadanalyse ist eine, auf ein volles rekursives Variablen-Modell angewandte multiple Regression.

Unter einem rekursivem Variablenmodell versteht man eine Kausalkette, die nur Kausalbeziehungen in eine gemeinsame Richtung beinhalten.

rekursiv

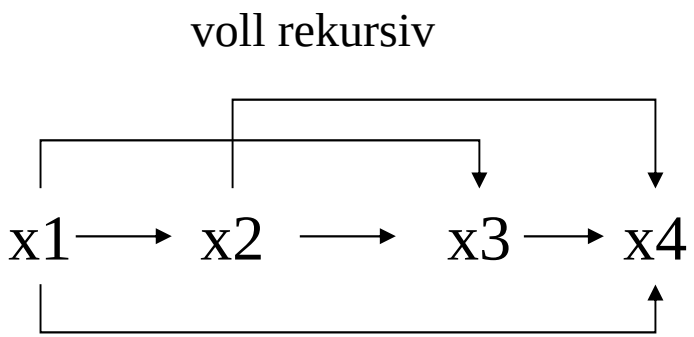


nicht rekursiv



volles rekursives Modell

Als volles rekursives Modell wird eine rekursive Kausalkette bezeichnet, in der alle nachgeordneten Variablen von allen vorangestellten Variablen kausal beeinflusst werden.

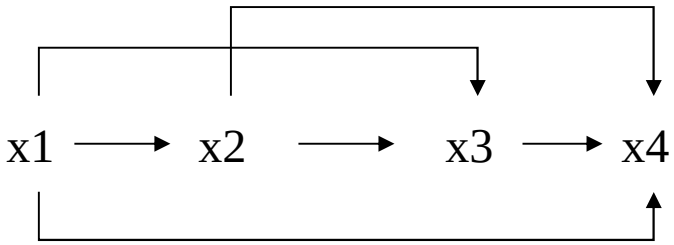


Das Verfahren der Pfadanalyse

Es wird eine Kausale Verknüpfung mehrerer Variablen vermutet und es kann eine kausale Reihenfolge dieser Variablen unterstellt werden.

Für die (theoretisch) angenommene Reihung wird zunächst ein volles rekursives Modell angenommen.

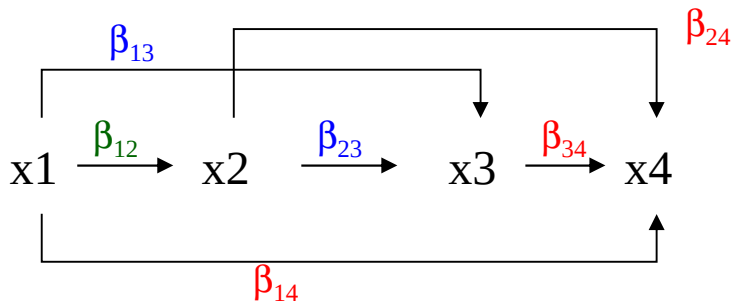
z.B.



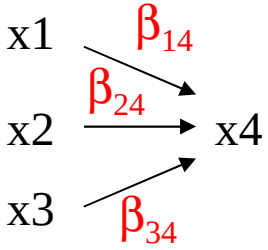
Das Verfahren der Pfadanalyse

Auf dieses angenommene volle rekursive Modell wird eine wiederholte multiple Regression angewandt.

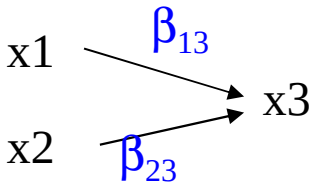
z.B. Modell:



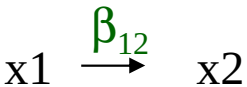
1. Regression:



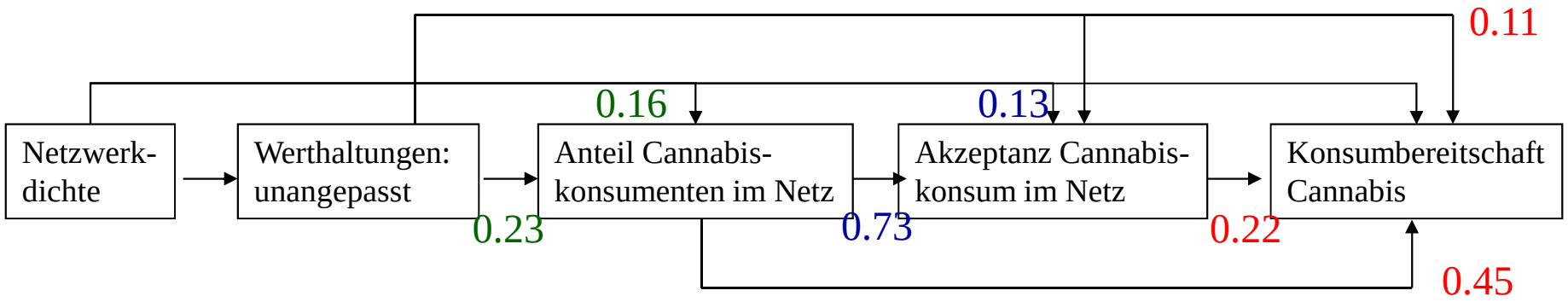
2. Regression:



3. Regression:



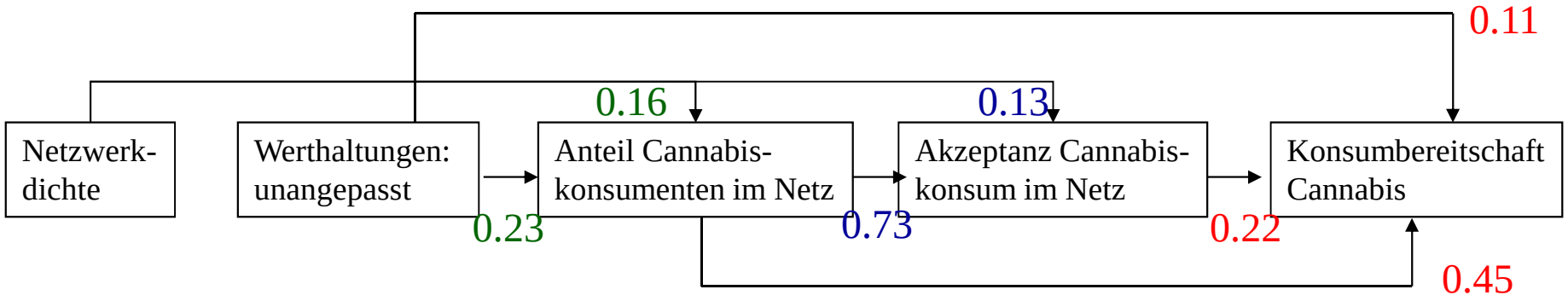
Das Verfahren der Pfadanalyse



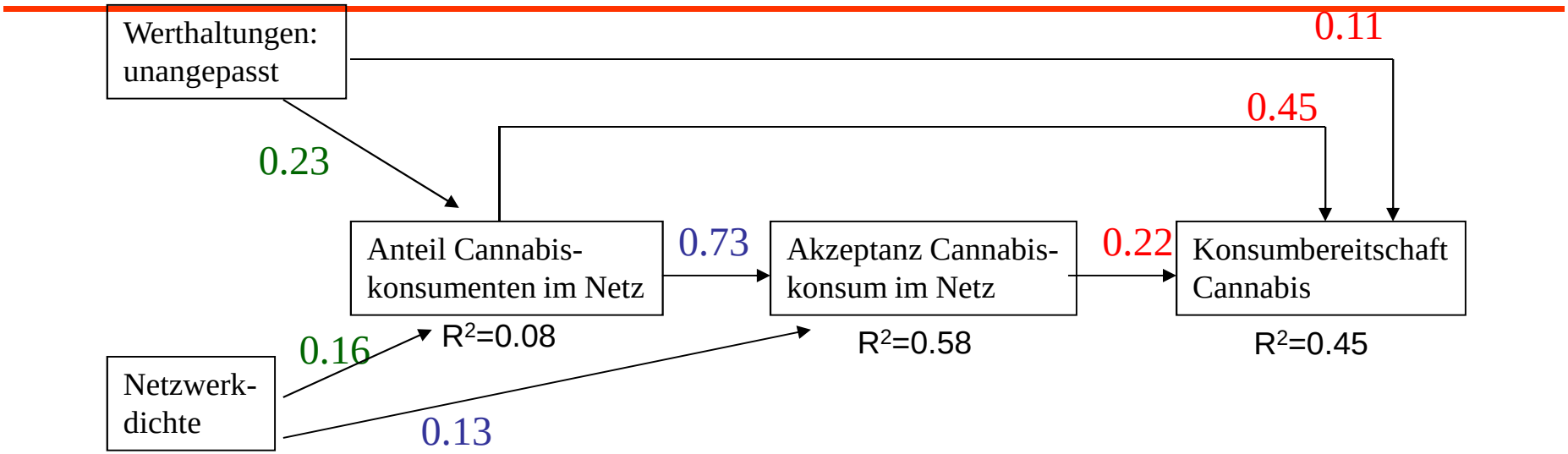
Variable	1	2	3	4
1 Dichte	0.03	**0.13**	**0.16**	0.04
2 Angepasst	**0.11**	0.00	**0.23**	abhängig
3 Anteil	**0.45**	**0.73**	abhängig	
4 Akzeptanz	**0.22**	abhängig		
5 Bereit	abhängig			
r^2	0.45	0.58	0.08	0.00

Das Verfahren der Pfadanalyse

Das volle rekursive Modell vereinfacht sich durch Pfade, für welche die β-Koeffizienten nicht signifikant > 0 sind.



Variable	1	2	3	4
1 Dichte	0.03	**0.13**	**0.16**	0.04
2 Angepasst	**0.11**	0.00	**0.23**	abhängig
3 Anteil	**0.45**	**0.73**	abhängig	
4 Akzeptanz	**0.22**	abhängig		
5 Bereit	abhängig			
r^2	0.45	0.58	0.08	0.00



Variable	1	2	3	4
1 Dichte	0.03	**0.13**	**0.16**	0.04
2 Angepasst	**0.11**	0.00	**0.23**	abhängig
3 Anteil	**0.45**	**0.73**	abhängig	
4 Akzeptanz	**0.22**	abhängig		
5 Bereit	abhängig			
r^2	0.45	0.58	0.08	0.00

Literaturempfehlung:

Wolf, Ch., Best, H. (Hrsg.)(2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesebaden, VS-Verlag. Kapitel 24.