Überblick: Multivariate Statistik

Zusammenhang / Unterschied /	Klassifikation / Reduktion
Test/Prognose	Exploration: Struktur
Varianzanalyse *Regressionsanalyse* Conjoint-Analyse Zeitreihenanalyse Diskriminanzanalyse	Faktorenanalyse *Clusteranalyse*

Klassische, häufigste Verfahren

Einfachster Fall: bivariate Regression

Zusammenhang zweier metrischer Variablen:

Unabhängige Variable (Prädiktor) → Abhängige Variable (Kriterium) (Modell)

Modellformulierung

Ist die postulierte Kausalität theoretisch begründet?

$$\boxed{y_i = b_0 + b_1 x_i + e_i}$$

(Funktion)

Ursachenanalyse

Gibt es einen (kausalen) Zusammenhang zwischen der unabhängigen und der abhängigen Variable? Wie eng ist dieser?

Wirkungsanalyse

Wie verändert sich die abhängige Variable bei einer Änderung der unabhängigen Variablen?

Prognose

Können die Messwerte der abhängigen Variable durch die Werte der unabhängigen Variable vorhergesagt werden?

Definition

zu schätzende Regressionskoeffizienten b_0 und b_1 (*ßs*)

Unabhängige Variable X

$$y_i = b_0 + b_1 x_i + e_i$$

Abhängige Variable Y

Fehlerwert (*Residuen*)

Voraussetzungen

Metrisches (intervallskaliertes) Niveau der Variablen

Linearität des Zusammenhangs

Linearität der Koeffizienten

Zufallsstichprobe

Für jeden Wert von x hat der Fehlerwert den Erwartungswert 0

Die Ausprägungen von x sind nicht konstant (Stichprobenvariation)

Für jeden Wert von x hat der Fehlerwert dieselbe Varianz (Homoskedaszitität)

(Gauss-Markov Annahmen)

Die Fehlerwerte hängen nicht voneinander ab (Unabhängigkeit) →

Autokorrelation bei Zeitreihen- und Paneldaten, Messwiederholungen

Die Fehlerwerte sind annähernd normalverteilt.

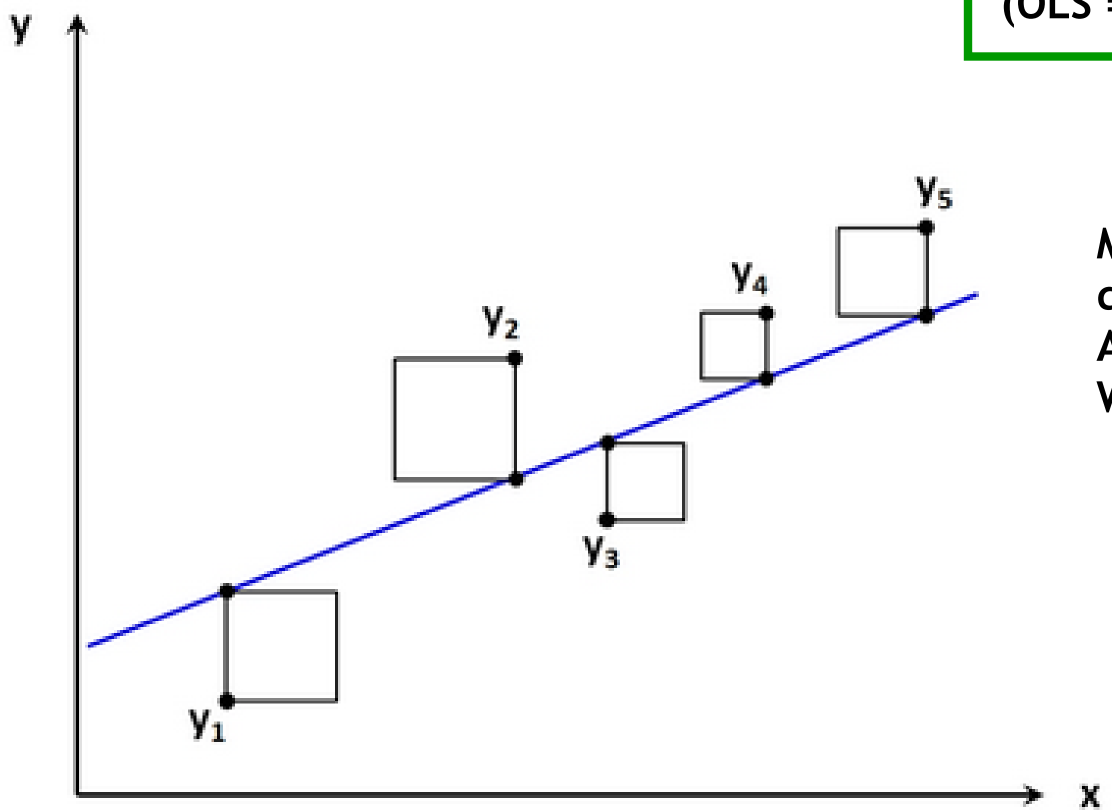
Regressionsgerade 1

Berechnung der Regressionsgeraden →

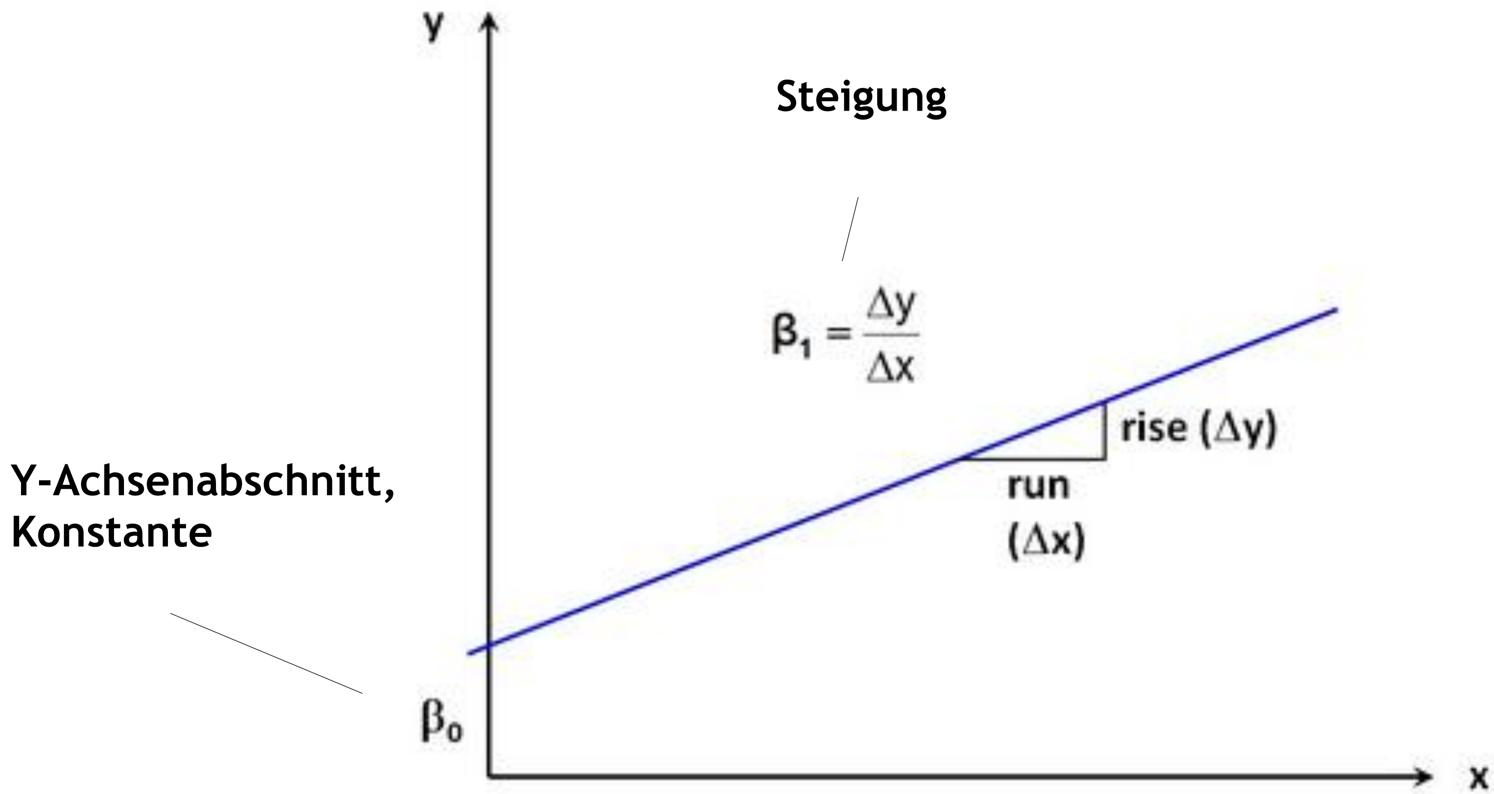
Methode der Kleinsten Quadrate (OLS = ORDINARY LEAST SQUARE)

↓

Minimierung der Summe der quadrierten vertikalen Abstände der beobachteten Werte zur Regressionsgeraden



Regressionskoeffizienten b_0 (β_0) und b_1 (β_1)



Der Fehlerterm e_i bezeichnet den Unterschied zwischen dem durch die Regressionsgerade vorhergesagten Wert y_i und dem tatsächlich gemessenen Wert y_i.

Anders gesagt: die Einflüsse auf y, die nicht auf x zurückgeführt werden können.

Prüfung der Voraussetzungen

metrisches Skalenniveau der Variablen: gegeben
Zufallsstichprobe: gegeben

(Untersuchungsanlage)

(1) Linearität

(2) Prüfung auf Erwartungswert 0 des Fehlerwerts e für x

(3) Prüfung der Varianz der Fehlerwerte (Heteroskedaszitität / Homoskedaszitität)

Streudiagramme: „Augenschein“

(4) Prüfung auf Unabhängigkeit der Fehlerwerte e

(5) Prüfung auf Normalverteilung der Fehlerwerte

Beispiel

Viele Menschen behaupten, dass bei ihnen erst der erste Schneefall Weihnachtsgefühle weckt. Eine Forschungsgruppe möchte untersuchen, ob Schneefall tatsächlich die Weihnachtsstimmung steigert. Eine bereits veröffentlichte Studie konnte zeigen, dass die Weihnachtsstimmung durch die Anzahl gekaufter Weihnachtsdekorationsartikel operationalisiert werden kann. Die Forschungsgruppe formuliert nun die folgende Forschungsfrage: Besteht ein Zusammenhang zwischen der Anzahl schneefallreicher Tage in der Vorweihnachtszeit und dem Umsatz (in Tausend Schweizer Franken) in Dekorationsgeschäften (n = 212)?

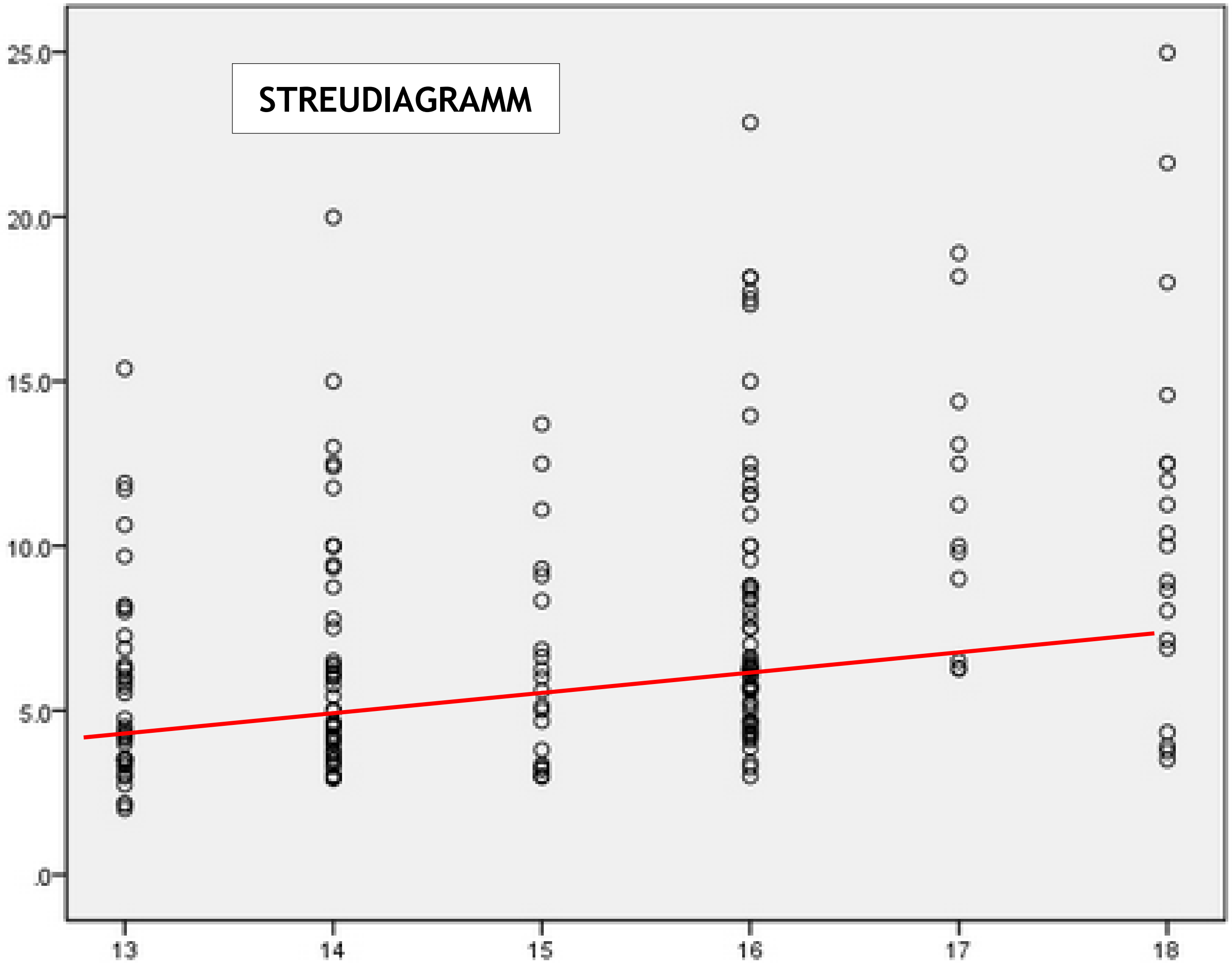
Unabhängige Variable „Schnee“

Abhängige Variable „Deko“

(1) Prüfung auf Linearität

Abhängige Var Y (Deko)

(Datensatz Schnee/Deko, analysiert mit SPSS)



Unabhängige Var X (Schnee)

Augenschein:

schwacher, positiver Zusammenhang

Was ist bei erkennbarer Nichtlinearität?

Transformation von y und/oder x

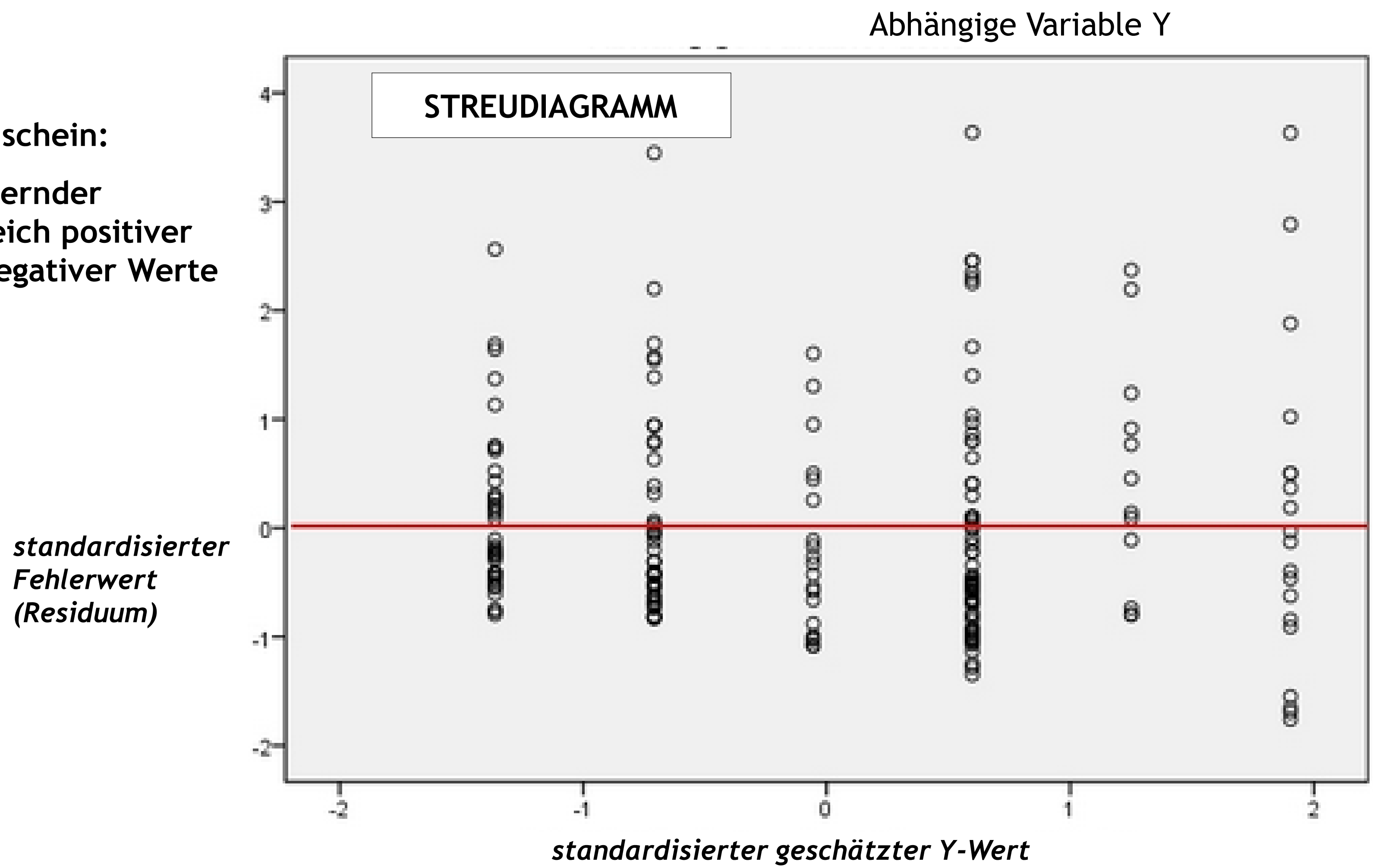
Logarithmierung, Quadrierung, Wurzel...

(2) Prüfung des Erwartungswerts des Fehlers e

Zur Erinnerung: Der Fehlerterm e_i bezeichnet den Unterschied zwischen dem durch die Regressionsgerade vorhergesagten Wert yi und dem tatsächlich gemessenen Wert yi.

Augenschein:

annähernder Ausgleich positiver und negativer Werte



Gauss-Markov-Annahme:

Für jeden Wert von x hat der Fehlerwert e dieselbe Varianz

- **Homoskedaszitität**
- **Heteroskedaszitität**

- **Streudiagramme: „Augenschein“**

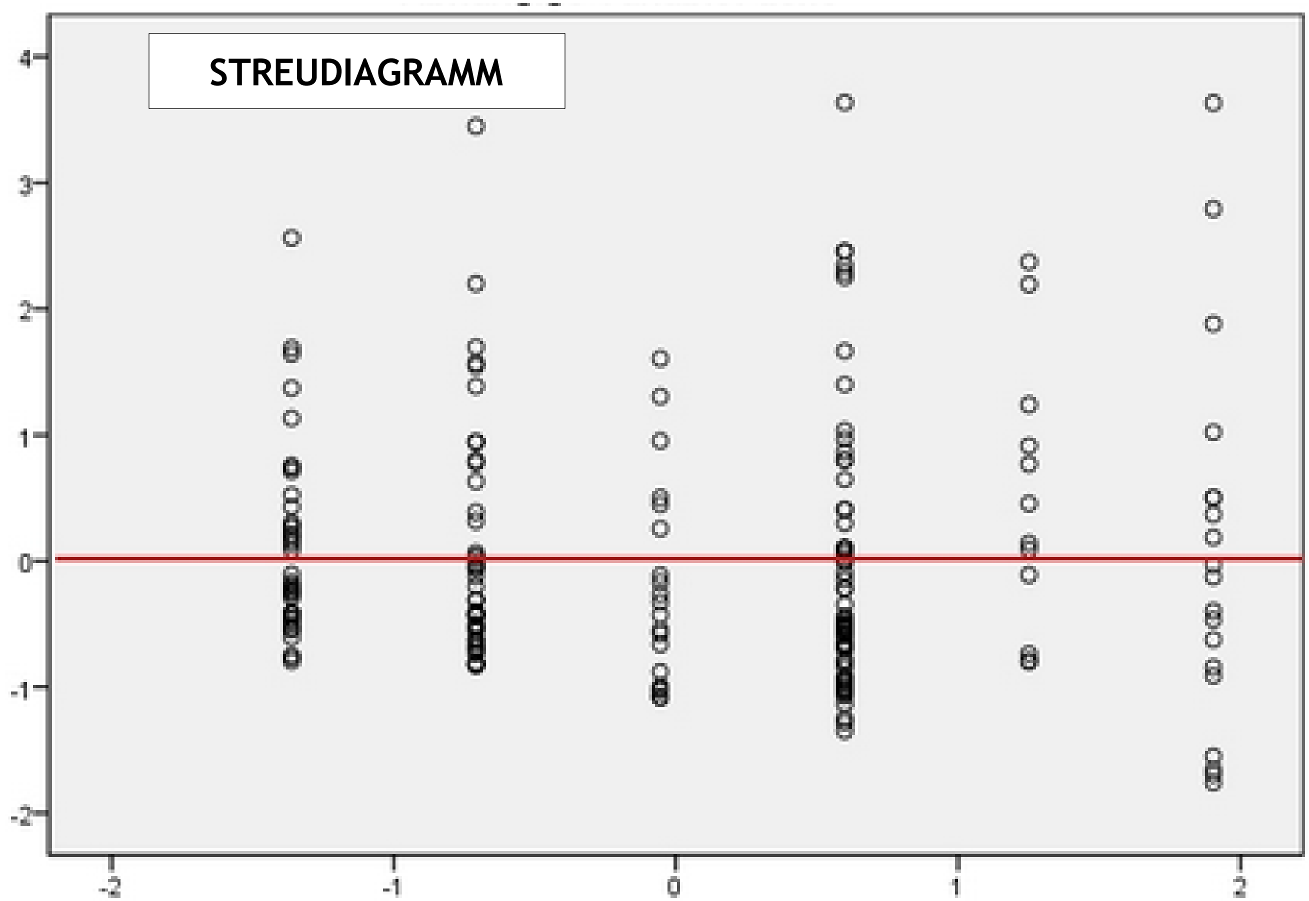
Außerdem:
Breusch-Pagan-Test / Cook-Weisberg-Test, White-Test

(3) Prüfung auf Homoskedaszitität

Augenschein:

kein „Muster“, also mutmaßlich Homoskedaszitität

Abhängige Variable Y



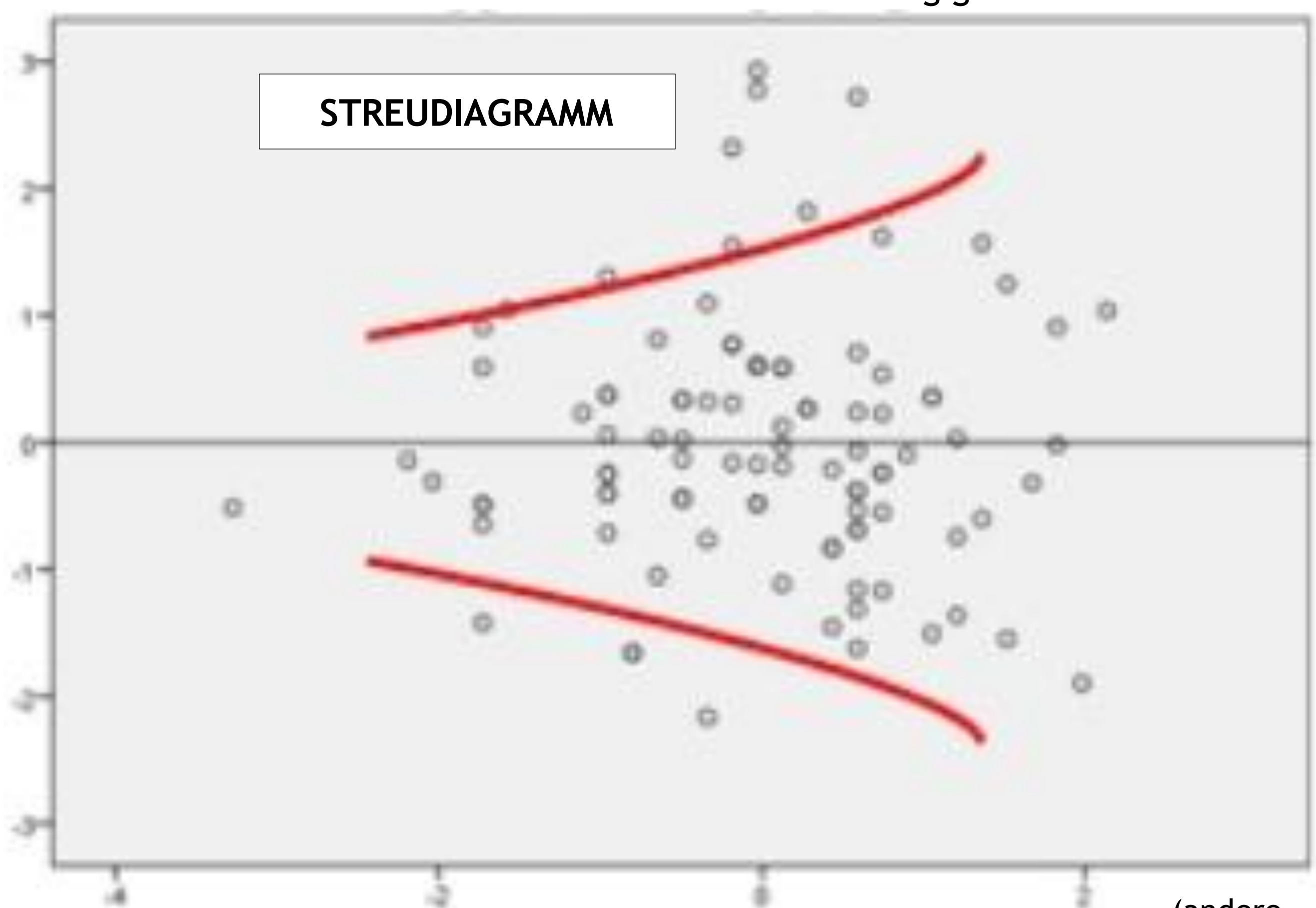
standardisierter Fehlerwert (Residuum)

standardisierter geschätzter Y-Wert

Wahrscheinliche Heteroskedaszitität =
Ungleichheit der Varianzen von e

Abhängige Variable Y

standardisierter
Fehlerwert
(Residuum)



standardisierter geschätzter Wert Y

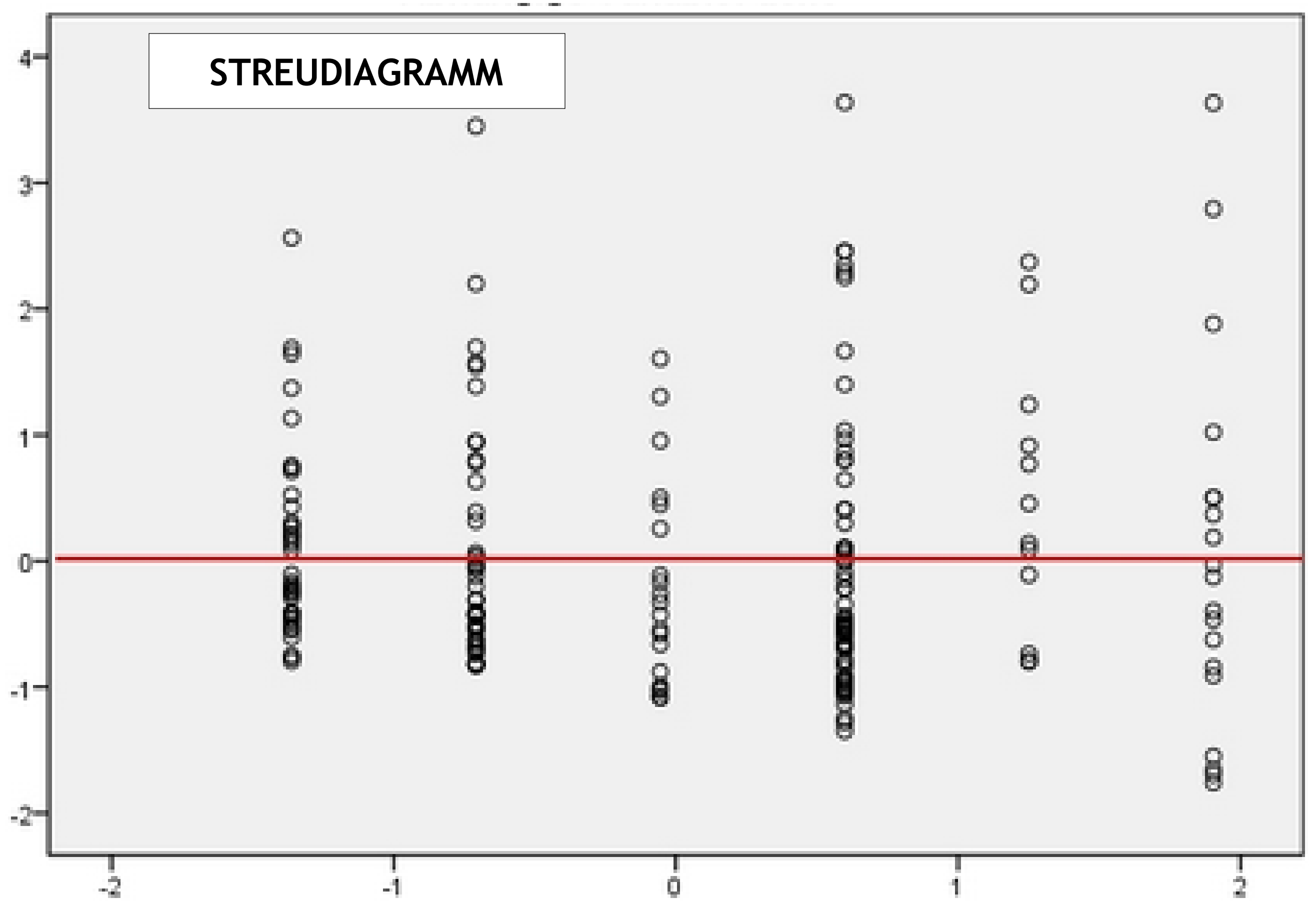
(andere
Beispieldaten)

(4) Prüfung auf Unabhängigkeit der Fehlerwerte

Augenschein:

kein „Muster“, also mutmaßlich keine Autokorrelation

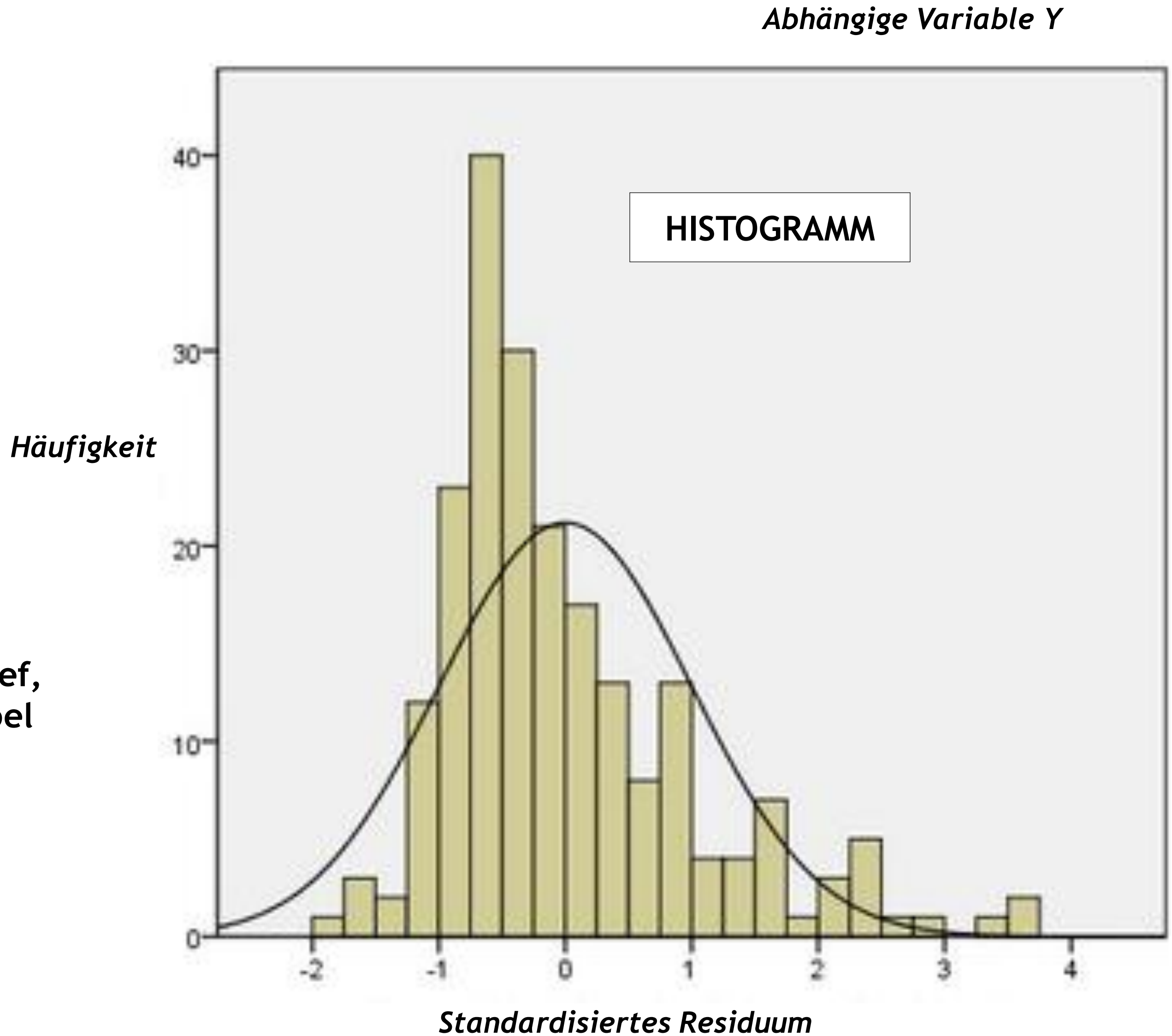
Abhängige Variable Y



standardisierter Fehlerwert (Residuum)

standardisierter geschätzter Y-Wert

(5) Prüfung auf Normalverteilung der Fehlerwerte



Augenschein:

Verteilung linksschief, aber noch akzeptabel

Fazit der Prüfung

Voraussetzungen nicht ideal, aber hinreichend gegeben.

Nächster Schritt: Signifikanzprüfung des Modells

Einfaktorielle Varianzanalyse (ANOVA) - F-Wert

Modellanalyse mit SPSS

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	563.166	1	563.166	35.451	.000[a]
	Nicht standardisierte Residuen	3335.999	210	15.886		
	Gesamt	3899.166	211			

$F(1,210) = 35.451, p = .000$

- **Das Modell als Ganzes ist signifikant**
- **Nächster Schritt: Schätzung der Koeffizienten**

Kleiner Exkurs: ANOVA

Definition

eine Verallgemeinerung des *t*-Tests für unabhängige Stichproben für Vergleich von mehr als zwei Gruppen (bzw. Stichproben)

Voraussetzungen

abhängige Variable intervallskaliert

unabhängige (Gruppen-/Treatment-) Variable nominal oder ordinal

Homogenität der Varianzen $\longrightarrow$ **T-Test-Anwendung**

Logik

Zerlegung der **Gesamtvarianz** der abhängigen Variable:

"Varianz innerhalb der Gruppen"
"Varianz zwischen den Gruppen"

sum of squares $_{total}$ — quadrierte Summe aller individuellen Abweichungen vom Mittelwert

=

sum of squares $_{zwischen}$ — Abweichungen zwischen Gesamtmittelwert und Gruppenmittelwerten

+

sum of squares $_{innerhalb}$ — individuelle Abweichungen vom jeweiligen Gruppenmittelwert

$$F = (SS_{zwischen} / df_{zwischen}) / (SS_{innerhalb} / df_{innerhalb})$$

Berechnung und Signifikanzprüfung

Verfahren: T - Test

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	-8.706	2.718		-3.204	.002
	Schnee	1.067	.179	.380	5.954	.000

Beide Koeffizienten (Konstante b_0 und b_1 sind hochsignifikant.

Interpretation

Konstante (b_0) = nicht 0 = Gerade nicht durch den Ursprung

b_1 = nicht 0 + signifikant = Einfluss X -> Y

$$y = -8.706 + 1.067x$$

→ Gerade/Funktion

b_1 = 1.067 = steigt x um eine Einheit, steigt y um 1.067

→ praktische Anwendung, Prognose

Frage

Wie gut passt das geschätzte Modell zu den Daten?

Prüfgröße

Bestimmtheitsmaß R^2

Welcher Anteil der Gesamtstreuung in der abhängigen Variable y wird durch x erklärt?

0 = keine Erklärungskraft

1 = perfekte Vorhersage von y

Je höher R^2, desto besser der „Fit“ des Modells

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.380[a]	.144	.140	3.9857

14,0 % der Gesamtstreuung von y wird durch x erklärt.

Frage

Wie bedeutsam ist ein gemessenes R^2?

Prüfgrößen

Mehrere Möglichkeiten zur Messung der Effektstärke
Die bekanntesten/gebräuchlichsten:

PEARSON-Korrelationskoeffizient r

COHEN- Effektstärkemaß f

$$\sqrt{\frac{R^2}{1-R^2}}$$

Beurteilung der Effektstärke nach COHEN

f = .10	**schwach**
f = .25	**mittel**
f = .40	**stark**

Im Beispiel:

$$\sqrt{\frac{0.14}{1-0.14}} = 0.40$$

Aussage

Die Anzahl schneereicher Tage (*schnee*) hat einen Einfluss darauf, wie viel Weihnachtsdekoration (*deko*) verkauft wird ($F(1, 210) = 35.451$, $p = .000$). Mit einem Tag mehr Schnee steigt der Umsatz an Weihnachtsdekoration um 1'067 Schweizer Franken. 14.0% der Streuung des Umsatzes an Weihnachtsdekoration wird durch die Anzahl schneereicher Tage erklärt, was nach Cohen einem starken Effekt entspricht.

Multiple Regression 1

Voraussetzungen

Intervallskalierte abhängige Variable Y, intervallskalierte oder Dummy-Variablen (0/1) X unabhängige Variablen

Linearität des Zusammenhangs

Linearität der Koeffizienten

Zufallsstichprobe

Für jeden Wert von x hat der Fehlerwert den Erwartungswert 0

Die Ausprägungen von x sind nicht konstant (Stichprobenvariation)

Für jeden Wert von x hat der Fehlerwert dieselbe Varianz (Homoskedaszitität)

Die Fehlerwerte hängen nicht voneinander ab (Unabhängigkeit)

Die Fehlerwerte sind annähernd normalverteilt.

Keine Multicollinearität, d.h. keine (starke) Korrelation zwischen den unabhängigen Variablen X_i

Wie im bivariaten Fall, aber mit einem Korrelationskoeffizienten je unabhängiger Variable (b_1 - b_k)

$$y = b_0 + b_1x_1 + b_2x_2 + ...b_kx_k + e_i$$

Konstante *Fehler*

Aussage

Wenn x_k um eine Einheit steigt, so verändert sich y um b_k Einheiten, gegeben alle anderen unabhängigen Variablen werden konstant gehalten.

Beispiel

Ein Club organisiert regelmäßig Konzerte. Um den Umsatz zu optimieren, möchten die Konzertveranstalter herausfinden, welche Faktoren zum Erfolg (Anzahl Besucher) eines Konzertes beitragen. Aus ihrer langjährigen Erfahrung wissen sie, dass der Erfolg unter anderem vom Ticketpreis (in Schweizer Franken), dem Werbeaufwand (in Schweizer Franken), sowie dem Erfolg der Band (Anzahl verkaufter CDs) abhängt. Dies möchte sie nun statistisch überprüfen, um künftig den Erfolg eines Konzertes im Voraus besser abschätzen zu können.

Der zu analysierende Datensatz enthält daher neben einer Identifikationsnummer des Anlasses (*ID*) die Besucherzahl des Anlasses (*Besuche*r), den Ticketpreis (*Preis*), den betriebenen Werbeaufwand (*Werbung*) und die Anzahl verkaufter CDs (*CD_Verkauf*).

Bestimmung der Aufnahmereihenfolge der unabhängigen Variablen x in das Modell

Reihenfolge spielt keine Rolle, wenn unabhängige Variablen nicht korrelieren → eher selten

Gleichzeitiger Einschluss

bei guter theoretischer Herleitung → Hypothesentest

Vorwärts-Selektion

Start: Variable X mit stärkster Korrelation mit Y

dann: jeweils Variable X mit stärkster partieller Korrelation mit Y

Stopp: wenn alle Variablen ins Modell aufgenommen sind oder sich R_2 (*Goodness of Fit*) nicht weiter signifikant erhöht

Rückwärts-Elimination

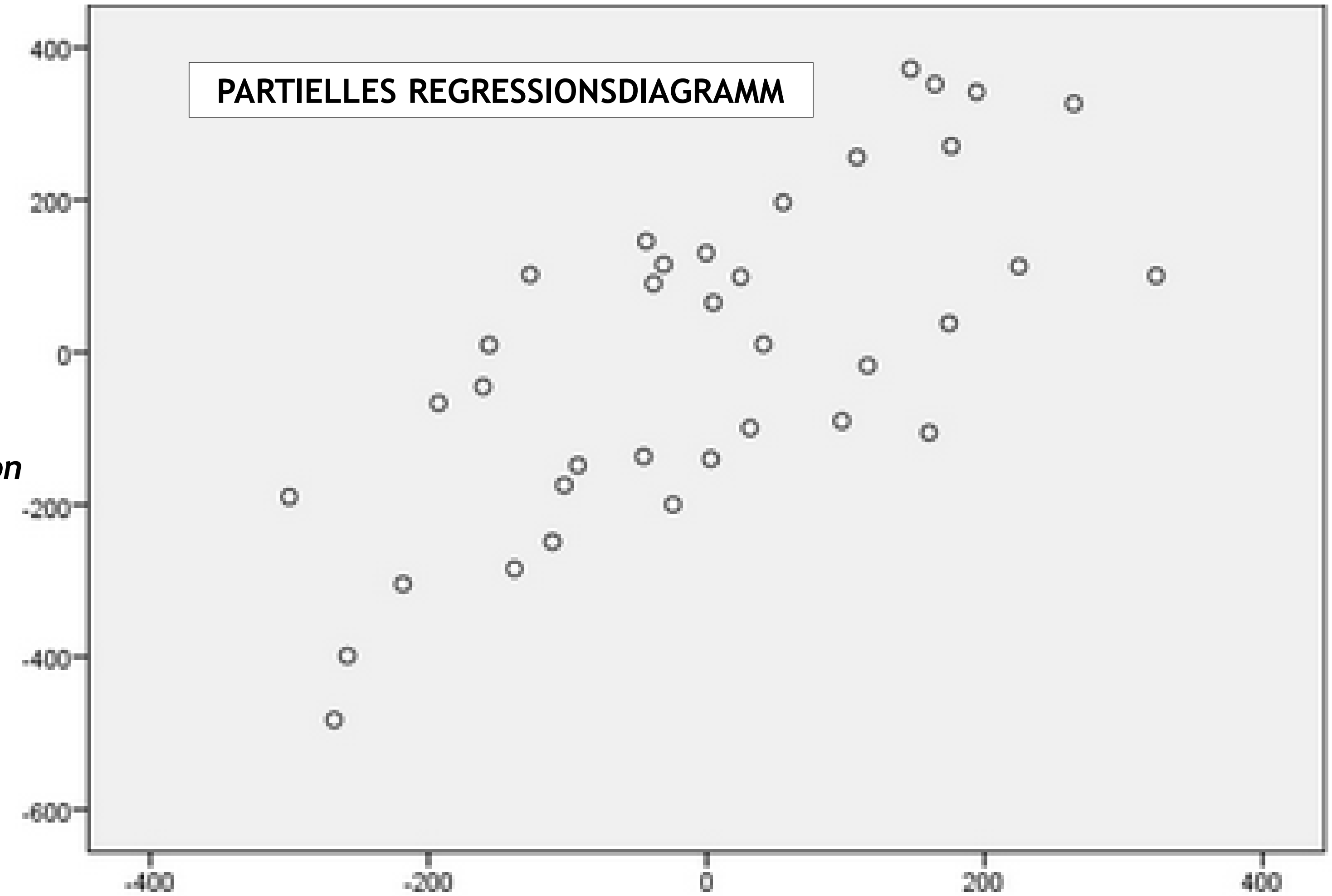
schrittweise die Variable X mit jeweils kleinster partieller Korrelation mit Y entfernen

Hierarchisch (blockwise)

Bündelung der unabhängigen Variablen in übergeordnete inhaltliche Blöcke

für alle Y X_n- Kombinationen: hier Y gegen X_1

Residuen: von allen anderen als X_1 NICHT erklärte Teil von Y (Besucher)



Residuen: von allen anderen X_n unabhängige Teil von X_1 (CD-Verkauf)

Multicollinearität und Autokorrelation

Die unabhängigen Variablen dürfen sich jeweils nicht als lineare Funktion der anderen unabhängigen darstellen lassen.

TOLERANZWERT

$$T_j = 1 - R_j^2$$

kleiner als .10

MULTICOLLINEARITÄT

VARIANZINFLATIONSFAKTOR

$$VIF_j = \frac{1}{1 - R_j^2}$$

größer als .10

Unabhängigkeit der Fehlerwerte

AUTOKORRELATION

DURBIN-WATSON

$$d := \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} (r_i)^2}$$

zwischen 0 und 4

0/4 = vollständige Autokorrelation

2 = keine Autokorrelation

Regressionskoeffizienten

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.	Kollinearitätsstatistik	
		Regressions koeffizientB	Standardfehler	Beta			Toleranz	VIF
1	(Konstante)	5091.213	1820.560		2.797	.009		
	Preis	-43.227	16.550	-.199	-2.612	.014	.990	1.011
	Werbung	.537	.056	.738	9.657	.000	.979	1.021
	CD_Verkauf	.965	.168	.439	5.759	.000	.983	1.017

a. Abhängige Variable: Besucher

Gleichung? Interpretation?

$$\text{Besucher} = 5091.21 - 43.23\,\text{Preis} + 0.54\,\text{Werbung} + 0.97 \cdot \text{CD\_Verkauf}$$

Goodness of Fit und Gesamtaussage

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	.904[a]	.817	.800	156.776	1.913

a. Einflußvariablen : (Konstante), CD_Verkauf, Preis, Werbung

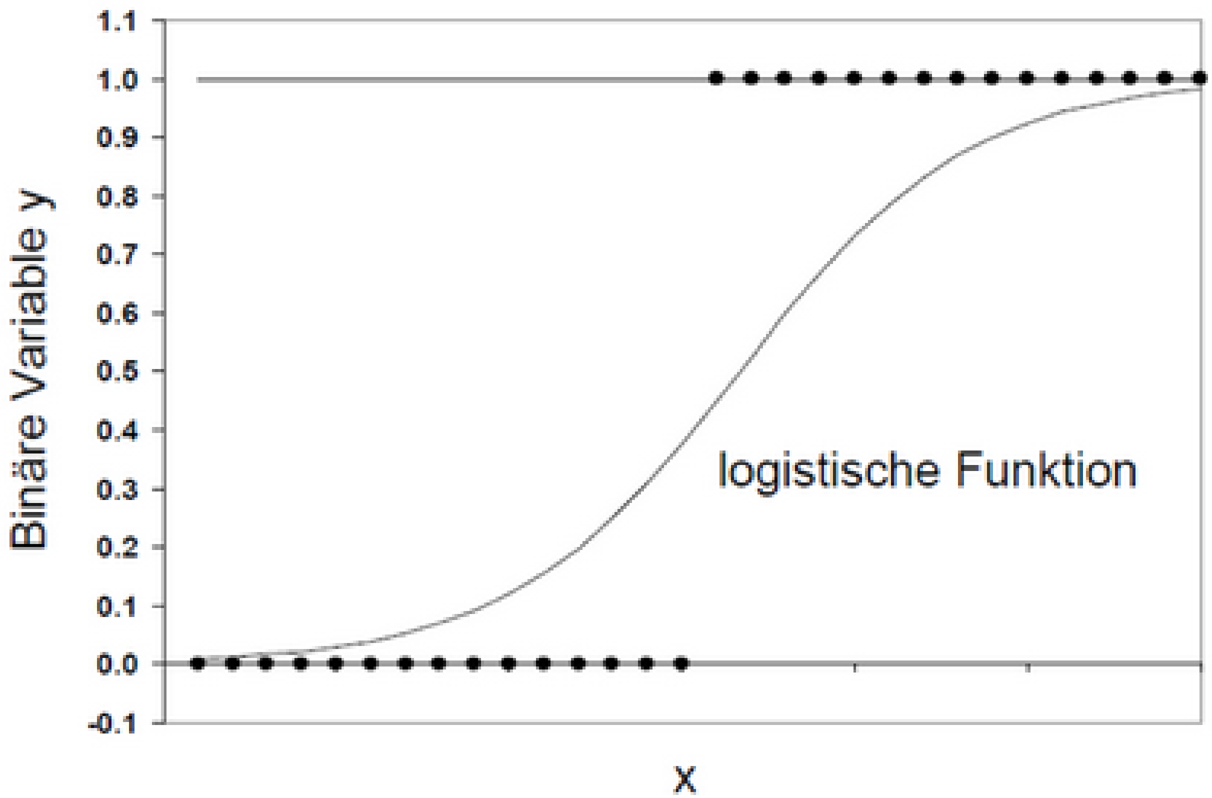
b. Abhängige Variable: Besucher

Eine multiple Regressionsanalyse zeigt, dass die Anzahl verkaufter Platten, der Ticketpreis sowie das Werbebudget einen Einfluss auf die Anzahl Konzertbesucher haben, $F(3,35) = 47.65$, $p = .000$, $n = 36$. Steigt der Preis der Konzertkarten um einen Schweizer Franken, so sinkt die Besucherzahl um durchschnittlich 43.23 Personen. Steigt das Werbebudget um einen Franken, nimmt die Anzahl Besucher um 0.54 Personen zu, und verkauft eine Band eine CD mehr, so nimmt die Besucherzahl um durchschnittlich 0.97 Personen zu. 80% der Streuung in der Anzahl Konzertbesuche wird durch die drei unabhängigen Variablen erklärt.

Zusammenhang einer abhängigen binären (1/0) und einer oder mehrer unabhängiger Variablen

→ **Wahrscheinlichkeit des Auftreten von 1**

→ **MAXIMUM-LIKLIHOOD-Schätzung (MLE)**



Logistische Funktion

Wahrscheinlichkeit, dass y = 1

$$\boxed{P(y=1)=\frac{1}{1+e^{-z}}}$$

Logit, d.h. Regressionsmodell

Eulersche Zahl

Logit $z=\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot x_3+\ldots+\beta_k\cdot x_k+\varepsilon$

$$P(y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot x_3+\ldots+\beta_k\cdot x_k+\varepsilon)}}$$

Beispiel

Eine Bank interessiert sich für Fakten, die mit der Wahrscheinlichkeit, dass jemand Aktien erwirbt, zusammenhängen. Sie beauftragt daher ein Marktforschungsinstitut, 700 Personen zu befragen. Es wird angenommen, dass der Entscheid für einen Aktienkauf vom Jahreseinkommen (in Tausend CHF), der Risikobereitschaft (Skala von 0 bis 25) sowie vom Interesse an der aktuellen Marktlage (Skala von 0 bis 45) beeinflusst wird.

Der zu analysierende Datensatz enthält daher neben einer Befragtennummer (*ID*) eine Variable zum Aktienkauf (*Aktienkauf*: 0 nein, 1 ja), das Jahreseinkommen (*Einkommen*), die Risikobereitschaft (*Risikobereitschaft*) und das Interesse an der aktuellen Marktlage (*Interesse*).

$$P(\text{Aktienkauf} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{Einkommen} + \beta_2 \cdot \text{Interesse} + \beta_3 \cdot \text{Risikobereitschaft})}}$$

Modellsignifikanz, Signifikanz der Koeffizienten, Goodness of Fit

Modell blockweiser Einschluss aller unabhängigen Variablen, Chi^2-Test

		Chi-Quadrat	df	Sig.
Schritt 1	Schritt	125.357	3	.000
	Block	125.357	3	.000
	Modell	125.357	3	.000

Variablen in der Gleichung

		Regressions koeffizientB	Standardfehler	Wald	df	Sig.	Exp(B)	95% Konfidenzintervall für EXP (B)	
								Unterer Wert	Oberer Wert
Schritt 1[a]	Einkommen	-.022	.006	14.651	1	.000	.979	.968	.990
	Risikobereitschaft	.348	.088	15.541	1	.000	1.416	1.191	1.683
	Interesse	.085	.018	23.036	1	.000	1.089	1.052	1.127
	Konstante	-1.668	.279	35.731	1	.000	.189		

a. In Schritt 1 eingegebene Variablen: Einkommen, Risikobereitschaft, Interesse.

Schritt	-2 Log-Likelihood	Cox & Snell R-Quadrat	Nagelkerkes R-Quadrat
1	679.007[a]	.164	.240

Vorhergesagte Werte und Wahrscheinlichkeiten

			Vorhergesagt		
			Aktienkauf		Prozentsatz der Richtigen
	Beobachtet		No	Yes	
Schritt 1	Aktienkauf	No	485	32	93.8
		Yes	135	48	26.2
	Gesamtprozentsatz				76.1

Gesamtaussage

Eine logistische Regressionsanalyse zeigt, dass sowohl das Modell als Ganzes (Chi-Quadrat(3) = 125.36, *p* = .000, *n* = 700) als auch die einzelnen Koeffizienten der Variablen signifikant sind. Steigen das Interesse an der Marktlage sowie die Risikobereitschaft um jeweils eine Einheit, so nimmt die relative Wahrscheinlichkeit eines Aktienkaufs um 8.9% beziehungsweise 41.6% zu. Steigt das Einkommen um 1'000 Franken, so sinkt die relative Wahrscheinlichkeit eines Aktienkaufs um 2.1%.

Das *R*-Quadrat nach Nagelkerke beträgt .24, was nach Cohen einem starken Effekt entspricht.