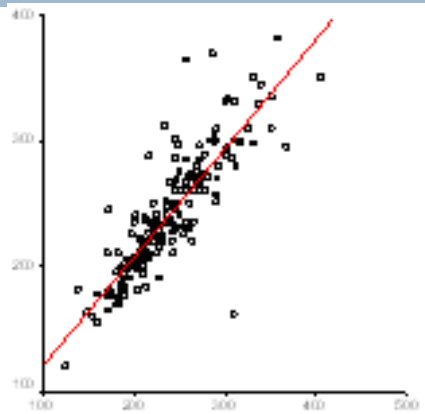
Regressionsanalyse



1. Einfache lineare Regression: Eine metrische AV, eine metrische oder dichotome UV

2. Multiple Regression: Eine metrische AV, mehrere metrische oder dichotome UVs

3. Ausblick: Logistische Regression: Eine kategoriale AV (2 oder mehr Ausprägungen), eine oder mehrere UVs unterschiedlicher Skalenniveaus

Einfache lineare Regression

- Vorhersage einer (metrischen) abhängigen Variable (AV) durch eine (metrische oder dichotome) unabhängige Variable (UV)

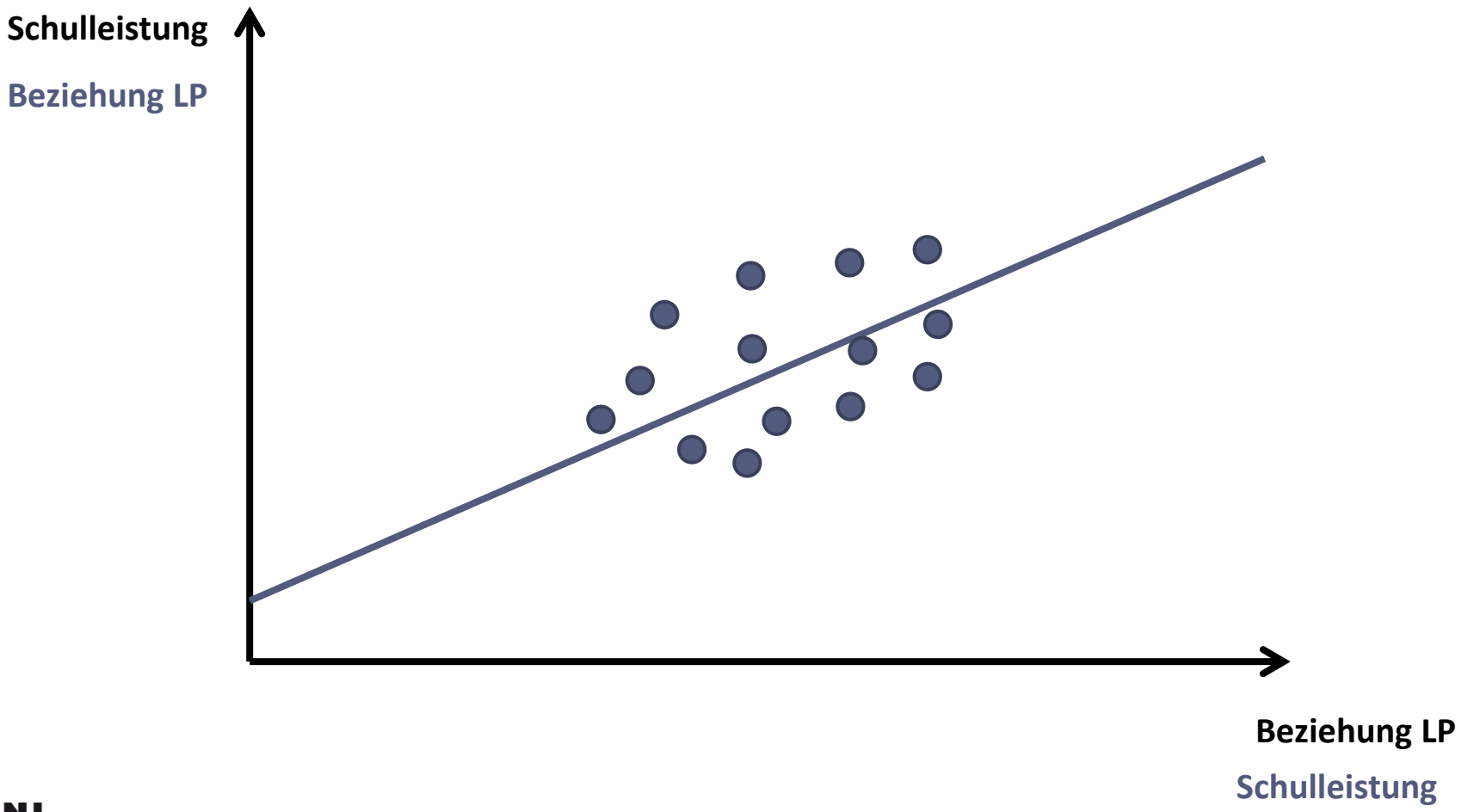
  Anders formuliert: Einfluss einer unabhängigen Variable auf eine abhängige Variable

- AV und UV können jedoch häufig ausgetauscht werden (was beeinflusst was?)

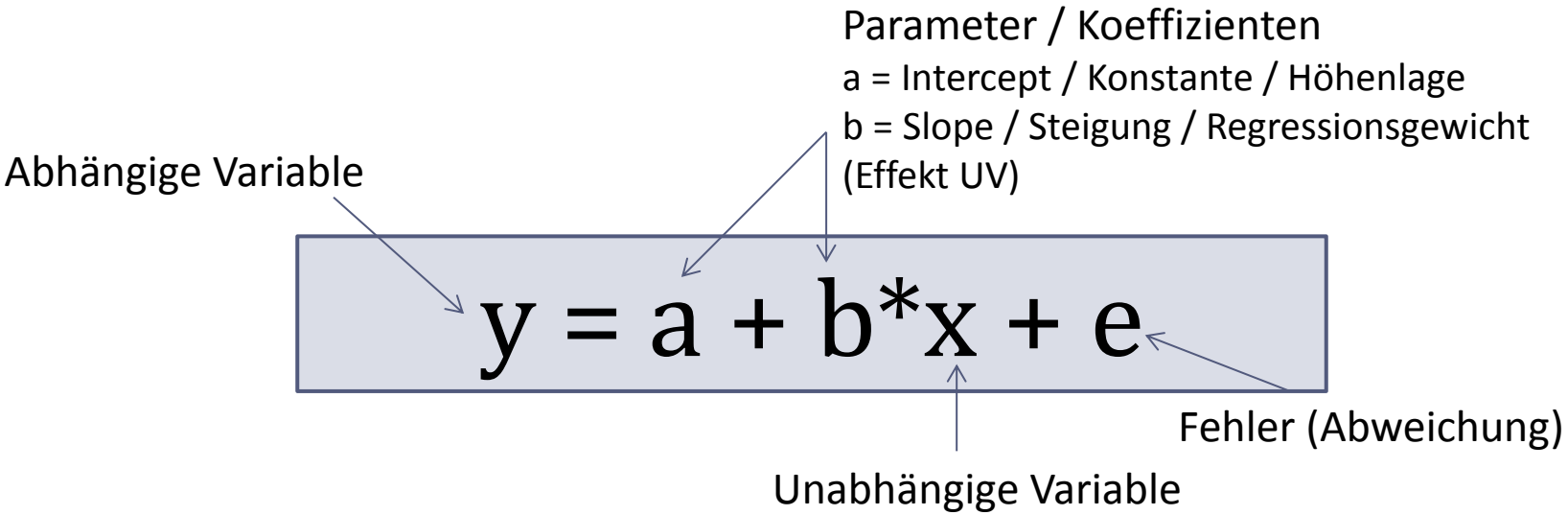
- Einflussrichtung nur dann klar definierbar, wenn ein Merkmal stabil ist oder wenn die beiden Variablen zeitlich versetzt gemessen wurden (-> Thema Messwiederholung)

- Deshalb: einfache lineare Regression = Korrelation -> gleiches Signifikanzergebnis und gleiche Effektstärke

Einfache lineare Regression

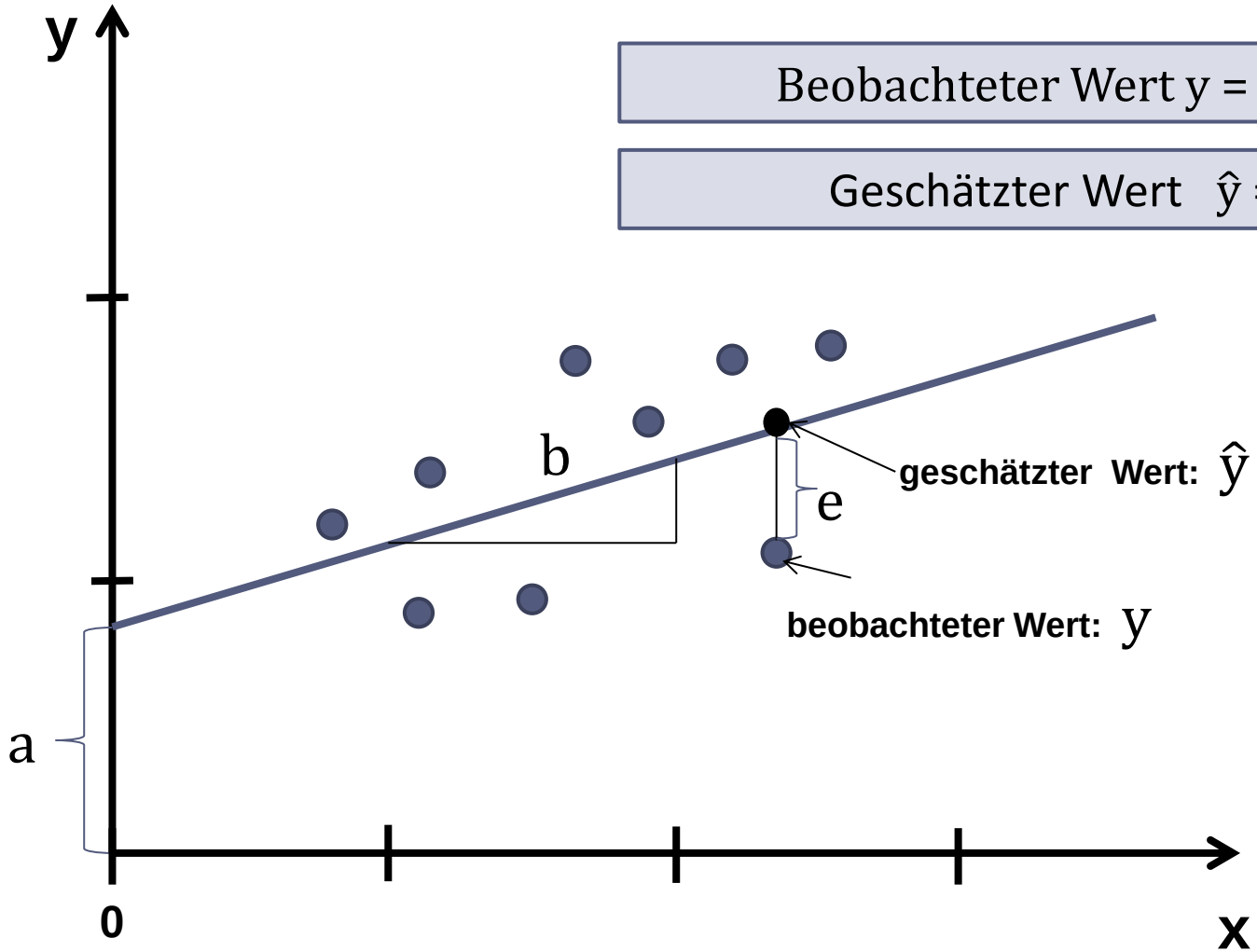


Die Regressionsgleichung (= allgemeines lineares Modell)



-> Alle behandelten statistischen Analyseverfahren sind Spezialfälle des allgemeinen linearen Modells (ALM)

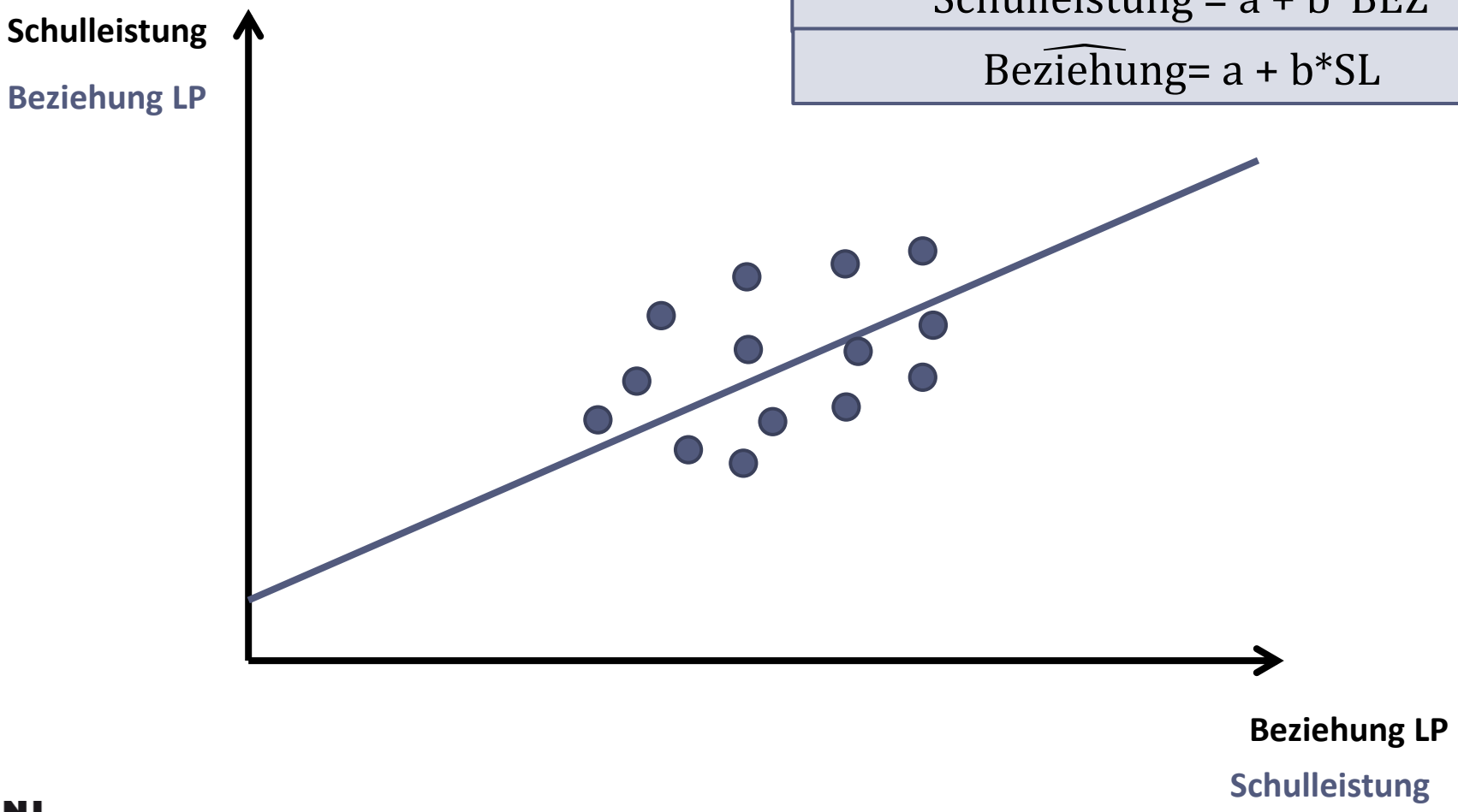
Die Regressionsgleichung (= allgemeines lineares Modell)



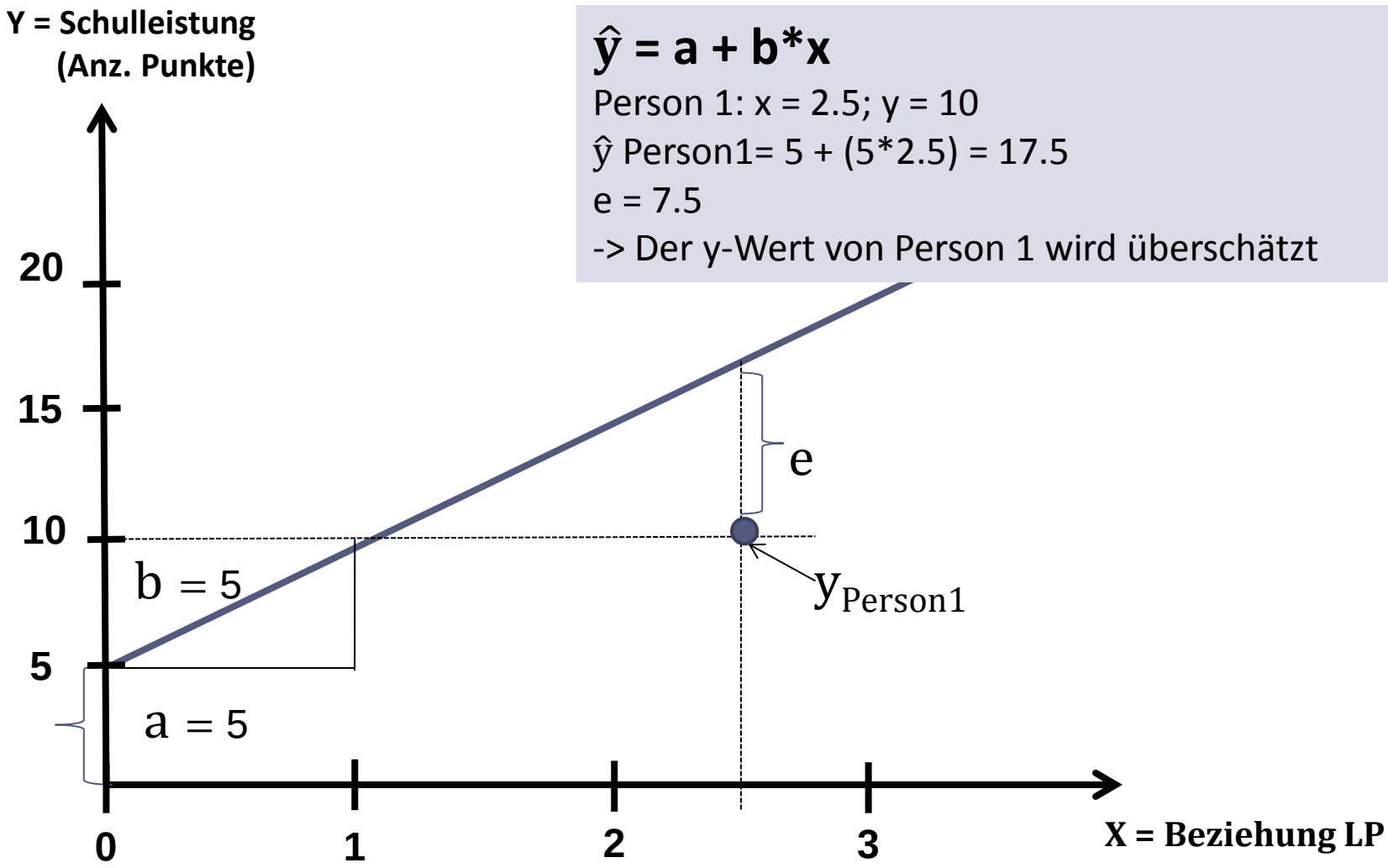
Die Regressionsgleichung (= allgemeines lineares Modell)

$$\widehat{\text{Schulleistung}} = a + b*BEZ$$

$$\widehat{\text{Beziehung}} = a + b*SL$$



Die Regressionsgleichung (= allgemeines lineares Modell)



$\hat{y} = a + b*x$

Person 1: x = 2.5; y = 10

$\hat{y}$ Person1= 5 + (5*2.5) = 17.5

e = 7.5

-> Der y-Wert von Person 1 wird überschätzt

Beispiel SPSS: Einfache lineare Regression

Was hat einen Einfluss auf körperliche Beschwerden?

Fragestellung : Haben psychische Probleme einen Einfluss auf körperliche Beschwerden?

H1 *Je mehr psychische Probleme jemand hat, desto mehr körperliche Beschwerden hat die Person*

-> Datensatz ALLBUS 2014

Variablen: PsychProb (0-4); Beschw (0-3.2); beides metrische Variablen, höhere Werte = mehr Probleme

Analysieren -> Regression -> linear -> AV (hier: Beschw.) und UV (hier: PsychProb) definieren

Beispiel SPSS: Einfache lineare Regression

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,533[a]	,284	,283	,70642

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	685,172	1	685,172	1373,016	,000[b]
	Residuum	1730,127	3467	,499		
	Gesamtsumme	2415,300	3468			

a. Abhängige Variable: Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

b. Prädiktoren: (Konstante), Psychische Probleme: MEAN(V230_2,V237_2,V238_2,V239_2)

Beispiel SPSS: Einfache lineare Regression

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	,400	,017		24,135	,000
	Psychische Probleme:	,560	,015	,533	37,054	,000

a. Abhängige Variable: Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

a = B0 = Intercept = Konstante = Mittelwert von y, wenn x = 0

B = Slope = Anstieg in y, wenn x um 1 Einheit ansteigt

$\hat{y} = a + b*x$

$\widehat{Beschw} = a + b*PsychProb$

$\widehat{Beschw} = .400 + .560*PsychProb$

Beispiel SPSS: Einfache lineare Regression: andere Einflussrichtung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	,400	,017		24,135	,000
	Psychische Probleme:	,560	,015	,533	37,054	,000

a. Abhängige Variable: Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	,338	,016		21,125	,000
	Körperliche Beschwerden:	,507	,014	,533	37,054	,000

a. Abhängige Variable: Psychische Probleme: MEAN(V230_2,V237_2,V238_2,V239_2)

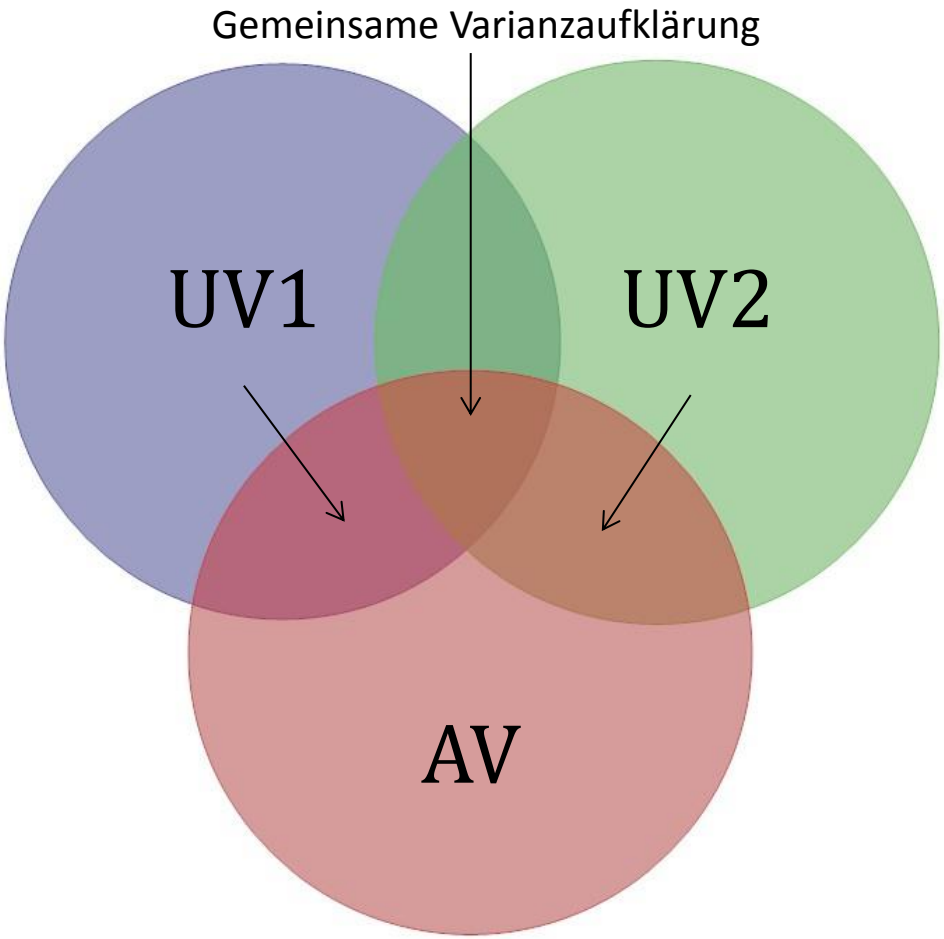
Beispiel SPSS: Einfache lineare Regression: Vergleich mit Korrelation

Korrelationen

		Psychische Probleme	Körperliche Beschwerden
Psychische Probleme	Pearson-Korrelation	1	,533**
	Sig. (2-seitig)		,000
	N	3469	3469
Körperliche Beschwerden	Pearson-Korrelation	,533**	1
	Sig. (2-seitig)	,000	
	N	3469	3469

**. Korrelation ist bei Niveau 0,01 signifikant (zweiseitig).

Multiple Regression



Multiple Regression

- Mehrere Variablen spielen eine Rolle bei der Vorhersage einer AV (sog. Kovariaten)

  -> $\hat{y} = a + b_1*x_1 + b_2*x_2 + b_3*x_3 + \ldots$

- Falls UVs unter sich korreliert sind, überschneiden sich ihre Varianzanteile
- Ein Teil der Erklärungskraft von UV1 für die AV wird von UV2 geteilt etc.
- Effekt von UV1 ist nun der «bereinigte» Effekt, welcher ausserhalb dieses Überschneidungsbereichs mit den anderen UVs liegt
- ≠ Korrelation von UV1 und AV

Vorteil gegenüber der einfachen linearen Regression und der Korrelation

- AV kann insgesamt besser vorhergesagt werden, weil meist von verschiedenen Variablen beeinflusst -> mehr Varianzaufklärung
- Störvariablen können zu Kontrollvariablen werden (z.B. individuelle Merkmale, kontextuelle Unterschiede etc.)
- Aufdecken von «Scheinkorrelationen»
- Man kann jedoch nie alles kontrollieren, deshalb: -> Berechtigt nicht zu kausalen Schlussfolgerungen in nicht-experimentellen Designs

Beispiel SPSS: Multiple Regression

Was hat einen Einfluss auf körperliche Beschwerden?

Fragestellung (1): Haben psychische Probleme einen Einfluss auf körperliche Beschwerden?

H1 (1) *Je mehr psychische Probleme jemand hat, desto mehr körperliche Beschwerden hat die Person*

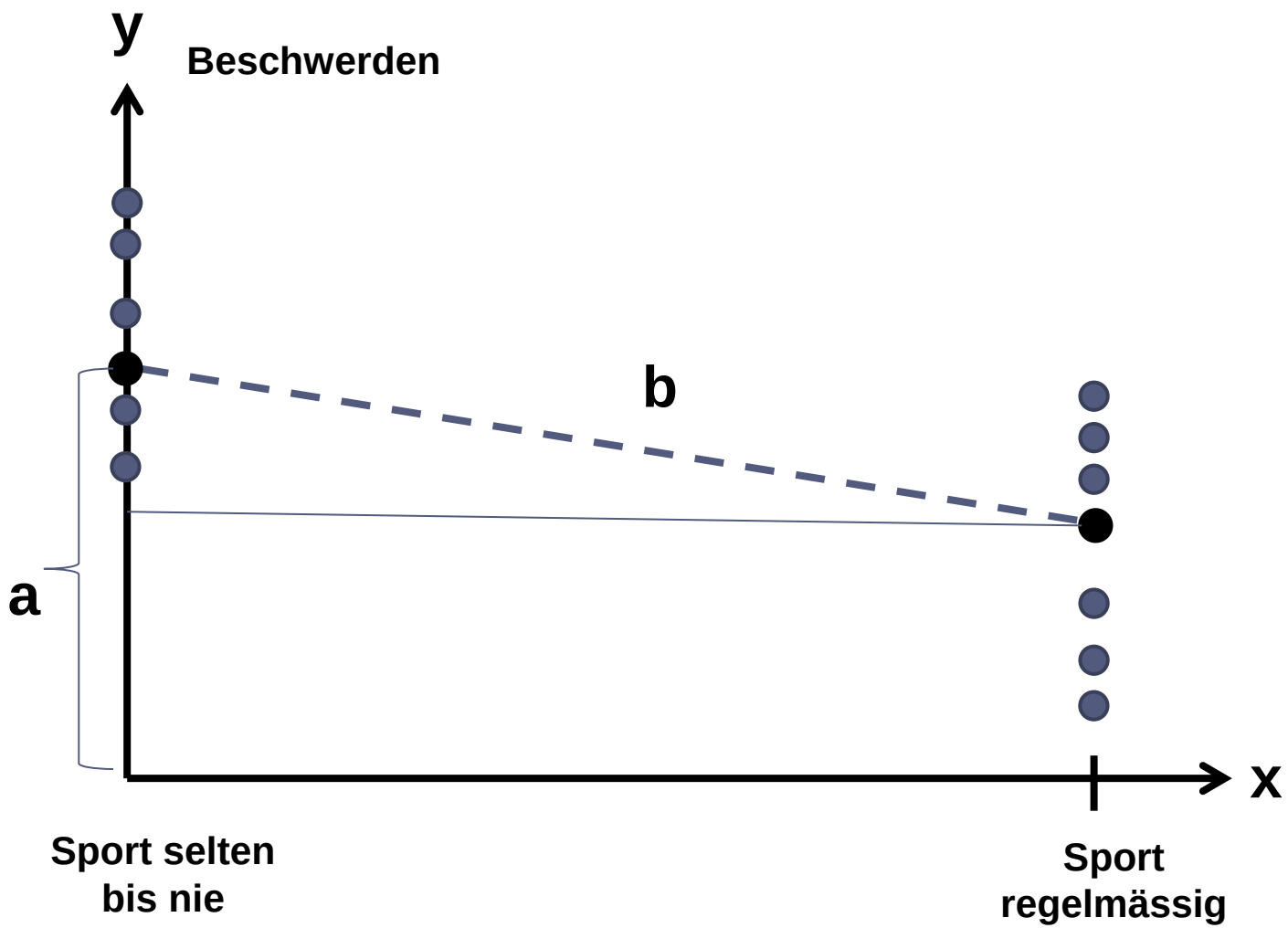
Fragestellung (2): Hat die sportliche Aktivität einen Einfluss auf körperliche Beschwerden? (= gibt es einen Unterschied zwischen Personen, die kaum Sport treiben und solchen, die regelmässig Sport treiben hinsichtlich körperlicher Beschwerden?)

H1 (2) *Personen, welche kaum Sport treiben, haben mehr körperliche Beschwerden als solche, welche regelmässig Sport treiben*

-> Datensatz ALLBUS 2014

Variablen:
- PsychProb (0-4), Beschw (0-3.2); beides metrische Variablen, höhere Werte = mehr Probleme
- Sport_01; 0=Sport selten bis nie; 1=Sport regelmässig

Unterschiede zwischen Gruppen als Regression?



Beispiel SPSS: Multiple Regression

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,554[a]	,307	,306	,69510

-> Zum Vergleich: Einfache lineare Regression ohne Variable Sport_01

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,533[a]	,284	,283	,70642

a. Prädiktoren: (Konstante), Psychische Probleme: MEAN(V230_2,V237_2,V238_2,V239_2)

Beispiel SPSS: Multiple Regression

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	740,983	2	370,492	766,798	,000[b]
	Residuum	1674,174	3465	,483		
	Gesamtsumme	2415,157	3467			

Beispiel SPSS: Multiple Regression

a = B0 = Intercept = Konstante = Mittelwert von y, wenn x1/x2 = 0

B1/B2 = Slopes = Anstieg (Abnahme) in y, wenn x um 1 Einheit ansteigt

Koeffizienten[a]

Modell	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
	Regressionskoeffizient B	Standardfehler	Beta		
(Konstante)	.547	.021		25.620	.000
PsychProb: Psychische Probleme	.541	.015	.515	36.116	.000
Sport_01: FREIZEIT: AKTIVE SPORTLICHE BETAETIGUNG dichotom	-.255	.024	-.153	-10.719	.000

a. Abhängige Variable: Beschw Körperliche Beschwerden

$y = a + b1*x1 + b2*x2$

$\widehat{Beschw} = a + b1*PsychProb - b2*Sport\_regelmässig$

$\widehat{Beschw} = .547 + .541*PsychProb - .255*Sport\_regelmässig$

Referenzkategorie = Sport_selten bis nie

Beispiel SPSS: Multiple Regression

Koeffizienten[a]

Modell	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
	Regressionskoeffizient B	Standardfehler	Beta		
(Konstante)	.547	.021		25.620	.000
PsychProb: Psychische Probleme	.541	.015	.515	36.116	.000
Sport_01: FREIZEIT: AKTIVE SPORTLICHE BETAETIGUNG dichotom	-.255	.024	-.153	-10.719	.000

a. Abhängige Variable: Beschw Körperliche Beschwerden

-> Zum Vergleich: Einfache lineare Regression ohne Variable Sport_01

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	,400	,017		24,135	,000
	Psychische Probleme:	,560	,015	,533	37,054	,000

Beispiel SPSS: Multiple Regression: Interpretation / Darstellung Ergebnisse

- Die erklärte Varianz des Gesamtmodells beträgt 30.6%. D.h. 30.6 % der Unterschiede zwischen den befragten Personen sind auf die Variablen «psychische Probleme» und «sportliche Aktivität» zurückzuführen
- Psychische Probleme haben einen signifikanten Einfluss auf körperliche Beschwerden ($p = .000$). D.h.: Je mehr psychische Probleme, desto mehr körperliche Beschwerden
- Sportliche Aktivität hat einen signifikanten Einfluss auf körperliche Beschwerden ($p = .000$). D.h. Personen, welche regelmässig Sport treiben, haben weniger körperliche Beschwerden als solche, welche selten bis nie Sport treiben
- Der Effekt der psychischen Probleme ($\beta = .52$) ist jedoch stärker als der Effekt der sportlichen Aktivität ($\beta = -.15$)

Tabelle 1

Regressionsanalyse zur Vorhersage von körperlichen Beschwerden durch psychische Probleme und sportliche Aktivität

Variable	*B*	*SE B*	β	*t*	*p*
Psychische Probleme	0.54	0.02	.52	36.12	.000
Sport regelmässig[a]	-0.26	0.02	-.15	-10.72	.000

Anmerkung. $R^2 = .31$ ($N = 3468$, $p < .01$).

[a]Referenzkategorie = Sport selten bis nie.

Nicol, A. A. M., & Pexman, P. M. (2010). *Presenting your findings. A practical guide for creating tables* (6. Aufl.). Washington, DC: American Psychological Association.

Übung Regressionsanalyse

Fragestellung 1: Haben psychische Probleme einen Einfluss auf die allgemeine Lebenszufriedenheit?

Fragestellung 2: Hat das Geschlecht einen Einfluss auf die allgemeine Lebenszufriedenheit

-> Datensatz ALLBUS 2014
Variablen:
Zufrieden (0-10); höhere Werte = mehr allgemeine Lebenszufriedenheit
PsychProb (0-4); höhere Werte = mehr Probleme
Geschl_01; männlich = 0; weiblich = 1

-> Formulieren Sie Hypothesen zu den Fragestellungen, führen Sie die Analyse durch und interpretieren Sie die Ergebnisse

Transformation von Variablen

- Zur sinnvollen Interpretation des Intercepts (Konstante)
- Kann UV keinen 0-Wert annehmen, ist die Interpretation des Intercepts nicht sinnvoll

-> Transformation: Differenz zum Nullwert, Zentrierung

Transformation von Variablen

Was hat einen Einfluss auf körperliche Beschwerden?

H1 (1) *Je mehr psychische Probleme jemand hat, desto mehr körperliche Beschwerden hat die Person*

H1 (2) *Personen, welche kaum Sport treiben, haben mehr körperliche Beschwerden als solche, welche regelmässig Sport treiben*

H1 (3) *Je älter eine Person ist, desto mehr körperliche Beschwerden hat sie*

-> Datensatz ALLBUS 2014

Variablen:
- PsychProb (0-4), Beschw (0-3.2); beides metrische Variablen, höhere Werte = mehr Probleme
- Sport_01; 0=Sport selten bis nie; 1=Sport regelmässig
- Alter (18-91)

Transformation von Variablen

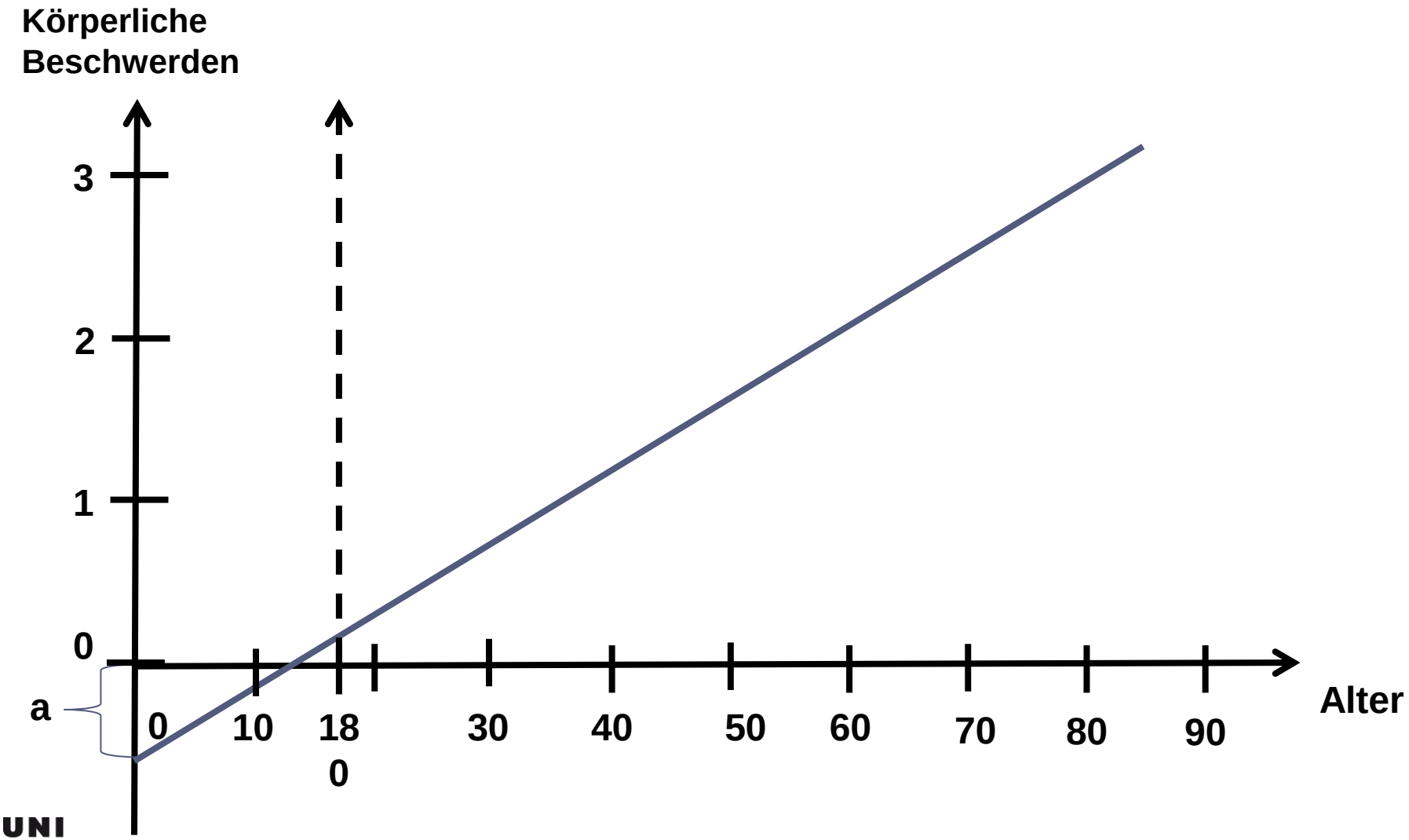
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	-.290	.038		-7.719	.000
	PsychProb	.535	.014	.509	39.043	.000
	Sport_01	-.177	.022	-.106	-8.054	.000
	Alter	.016	.001	.340	26.011	.000

a. Abhängige Variable: Beschw Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

Mittelwert körperlicher Beschwerden, wenn psychische Probleme = 0, Sport = 0 (selten bis nie) und Alter = 0

Transformation von Variablen

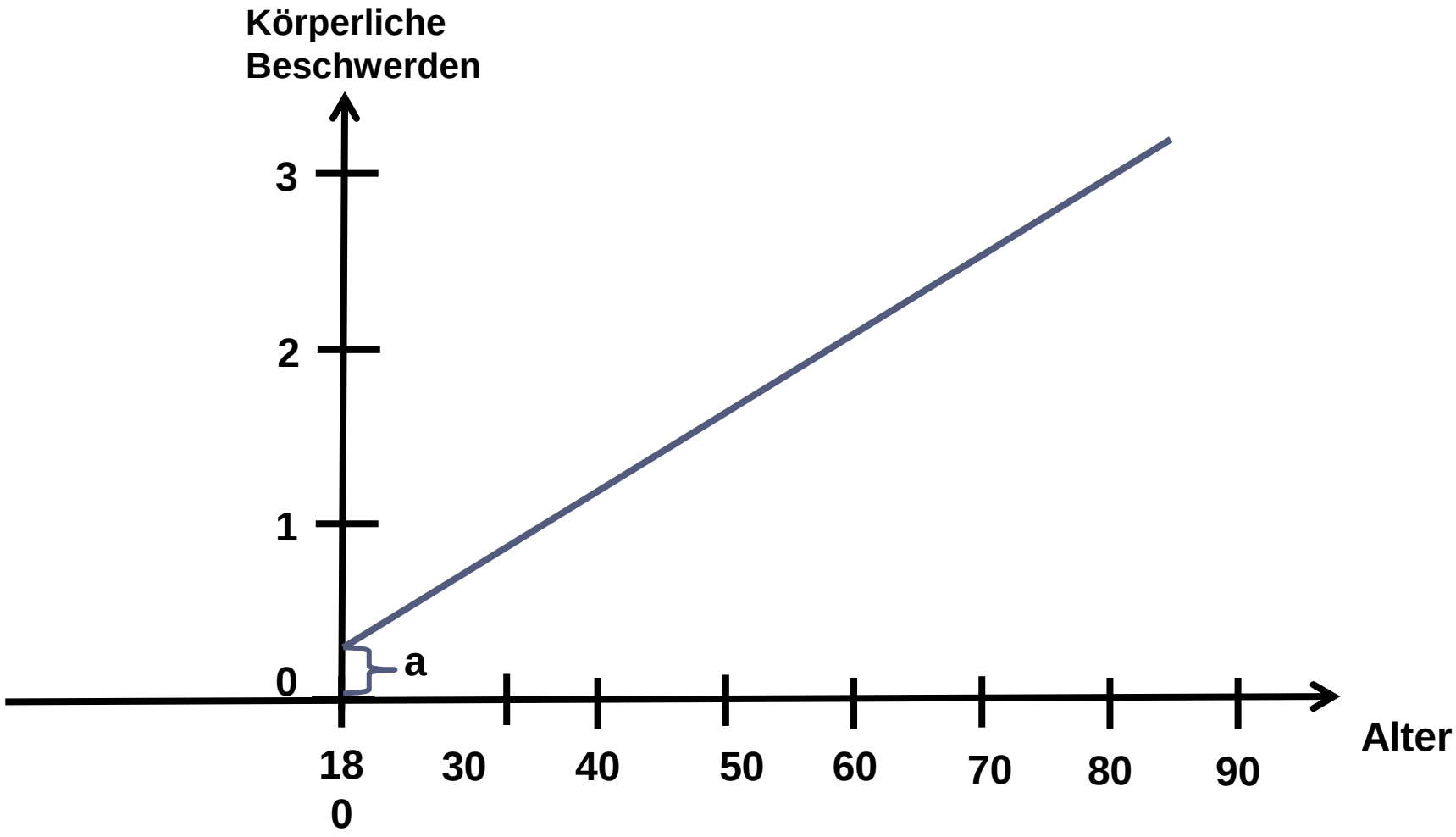


UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Transformation von Variablen: Differenz zum Nullpunkt

- Beispiel Variable Alter: Beginnt bei 18 -> also: von allen Werten 18 abziehen
- Der neue Mittelwert ist 31.44 (vorher: 49.44)
- Der Wert 0 der UV bezieht sich nun auf das Alter 18 -> also: Intercept = Mittelwert der AV, wenn das Alter 18 Jahre beträgt

Transformation von Variablen: Differenz zum Nullpunkt



Transformation von Variablen: Differenz zum Nullpunkt

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	.001	.029		.043	.966
	PsychProb	.535	.014	.509	39.043	.000
	Sport_01	-.177	.022	-.106	-8.054	.000
	Alter18 Alter: Wert 18 abgezogen	.016	.001	.340	26.011	.000

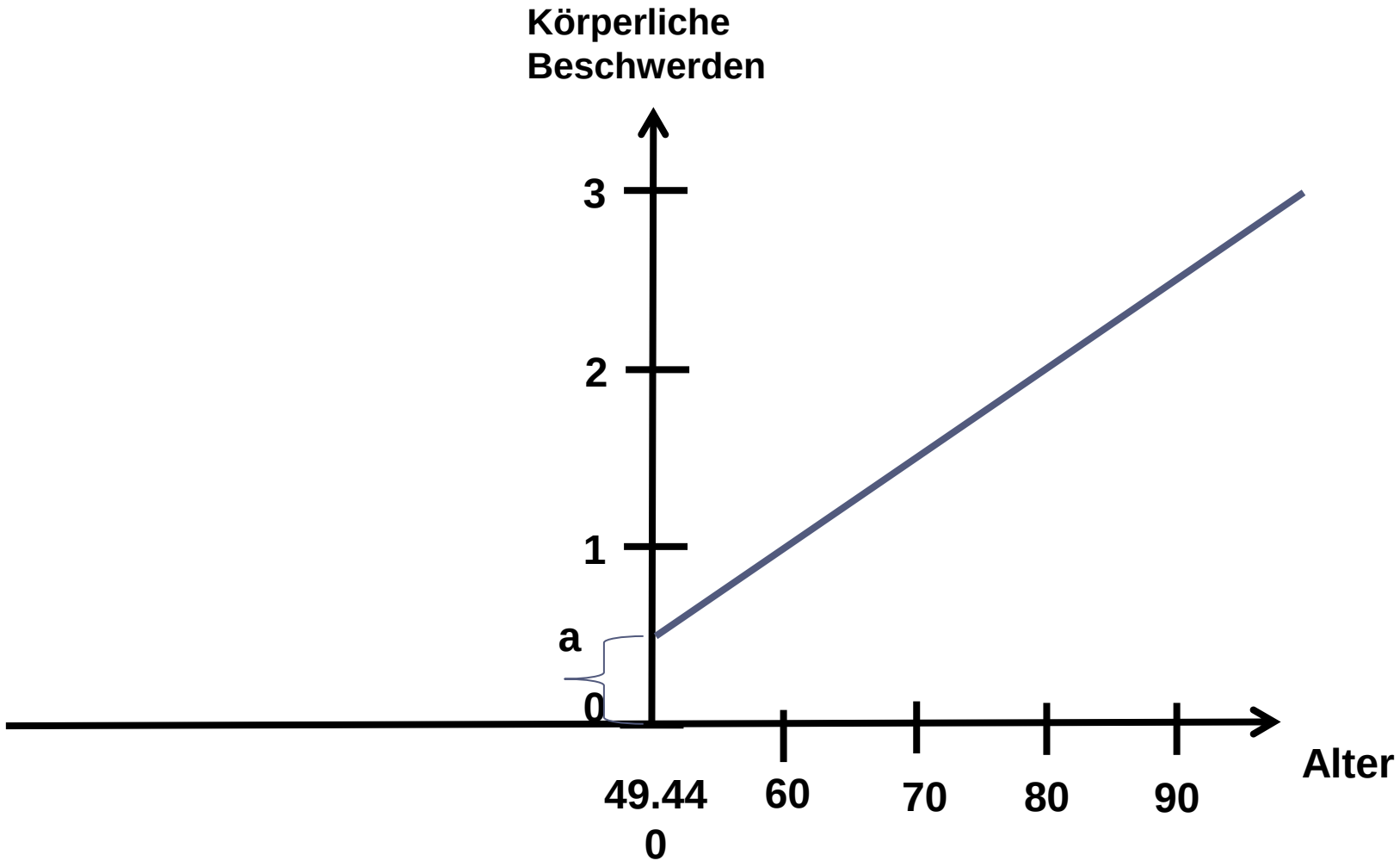
a. Abhängige Variable: Beschw Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

Mittelwert körperlicher Beschwerden, wenn psychische Probleme = 0, Sport = 0 (selten bis nie) und Alter = 18 J.

Transformation von Variablen: Zentrierung

- Zentrierung: $x_i - \bar{x}$ = Wert Person i – Gesamtmittelwert (deshalb auch: «grand-mean-centering»)
- Beispiel Variable Alter: $\bar{x}$ = 49.44 -> von allen Werten wird der Wert 49.44 abgezogen
- Der neue Mittelwert ist 0 (positive und negativen Abweichungen heben sich auf)
- Der Wert 0 der UV bezieht sich nun auf das Alter 49.44 -> also: Intercept = Mittelwert der AV, wenn das Alter 49.44 Jahre beträgt

Transformation von Variablen: Zentrierung



Transformation von Variablen: Zentrierung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	.510	.020		26.068	.000
	PsychProb	.535	.014	.509	39.043	.000
	Sport_01	-.177	.022	-.106	-8.054	.000
	AlterZent Alter zentriert um Mittelwert M=49.44	.016	.001	.340	26.011	.000

a. Abhängige Variable: Beschw Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

Mittelwert körperlicher Beschwerden, wenn psychische Probleme = 0, Sport = 0 (selten bis nie) und Alter = 49.44

Transformation von Variablen in SPSS

Differenz zum Nullpunkt oder Zentrierung:
Transformieren -> Variable berechnen -> Zielvariable beschriften -> Zielvariable = Numerischer Ausdruck: ursprüngliche Variable (hier: Alter) minus den entsprechenden Wert (hier: 18 resp. 49.44) -> neue Variable erscheint zuunterst im Datensatz

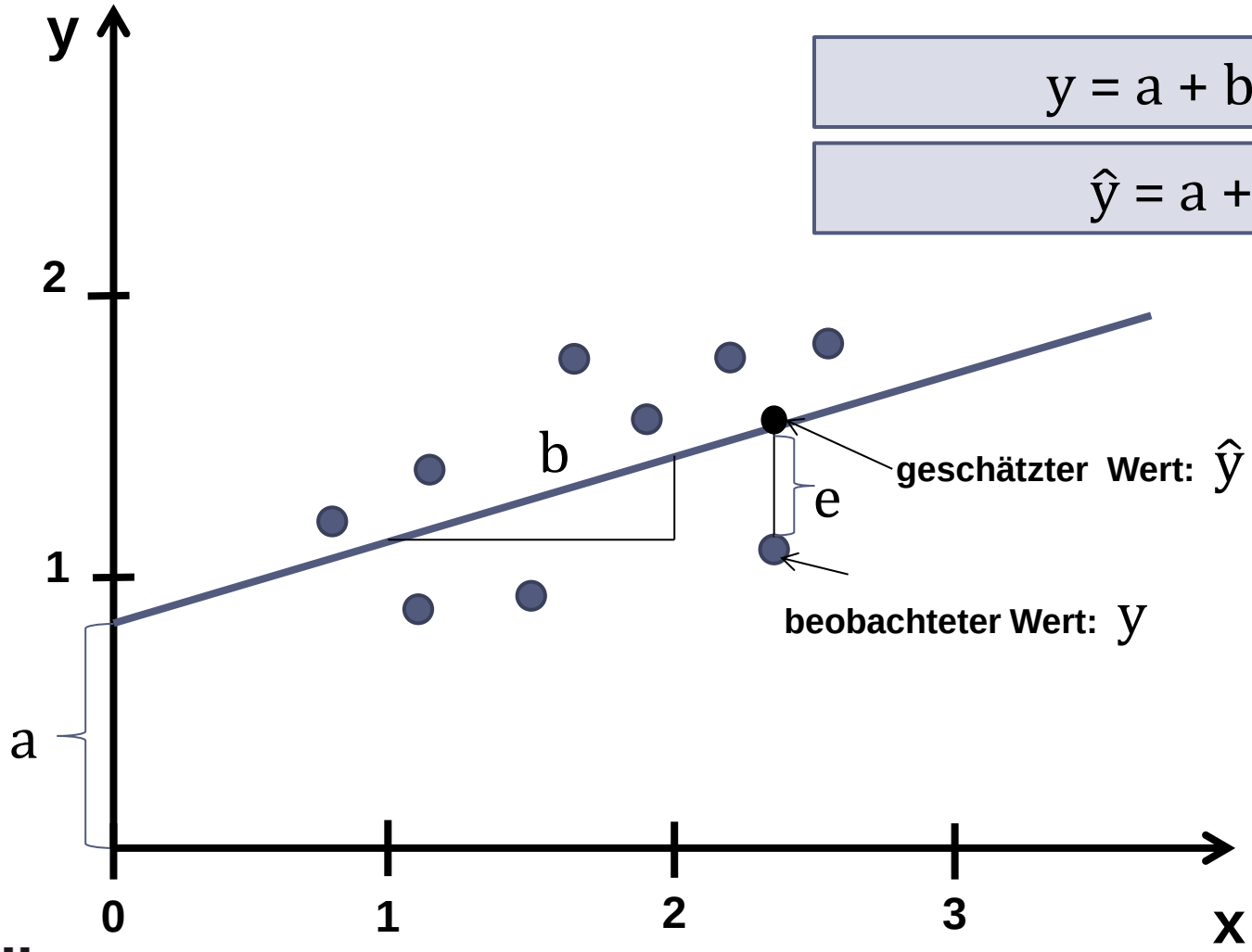
Transformation von Variablen: Schlussfolgerungen

- Sinnvoll für die Interpretation des konstanten Terms
- Die besprochenen Verfahren sind lineare Transformationen, d.h. es hat keinen Einfluss auf die Verteilung der Daten
- Keinen Einfluss auf die Effekte der UVs
- Für die deskriptiven Statistiken, welche den Hypothesentests vorangehen, ist eine Transformation meist nicht sinnvoll

Modellprämissen der Regression: Übersicht

Prämisse	Prämissenverletzung	Konsequenzen
Normalverteilung der Residuen	Nicht-normalverteilte Residuen	Ungültige Signifikanztests (bei kleinen Stichproben < 40)
Linearität	Nichtlinearität	Verzerrung der Schätzwerte (B-Koeffizienten)
Homoskedastizität (Homogenität der Varianzen)	Heteroskedastizität (Inhomogenität der Varianzen)	Verzerrung der Standardfehler -> ungültige Signifikanztests
Unabhängigkeit der Residuen	Autokorrelation der Residuen	Verzerrung der Standardfehler -> ungültige Signifikanztests
Unabhängigkeit der UVs untereinander	Multikollinearität	Verminderte Präzision der Schätzwerte (B-Koeffizienten); heben sich gegenseitig auf

Was sind Residuen?



$$y = a + b*x + e$$

$$\hat{y} = a + b*x$$

Nicht-normalverteilte Residuen

-> Folge: Ungültige Signifikanztests

1. Visuelle Inspektion: Histogramm, Normalverteilungsdiagramm

   Analysieren -> Regression -> Linear: AV (hier: Beschw)und UVs (hier: PsychProb; Sport_01; Alter18) definieren

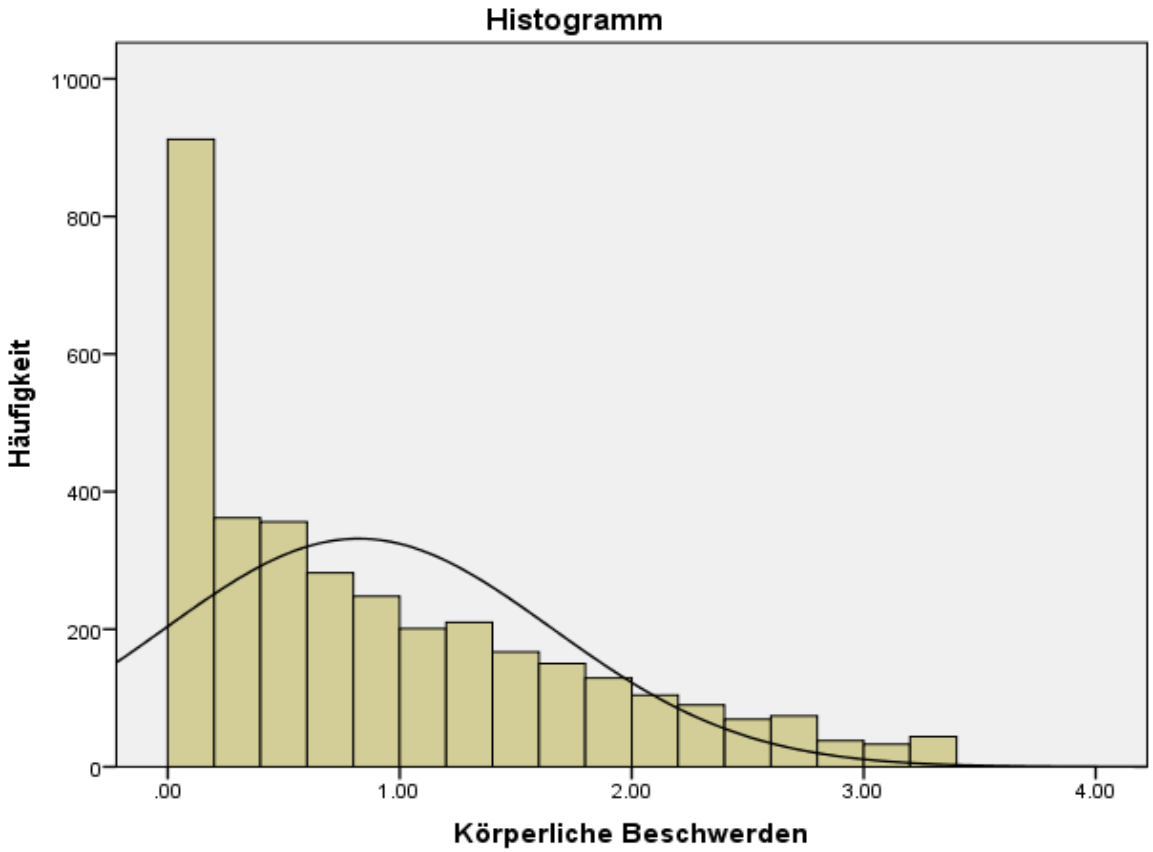
   -> Diagramme: Histogramm und Normalverteilungsdiagramm

2. Test auf Normalverteilung: Kolmogorov-Smirnov und Shapiro-Wilk (mehr Testpower)

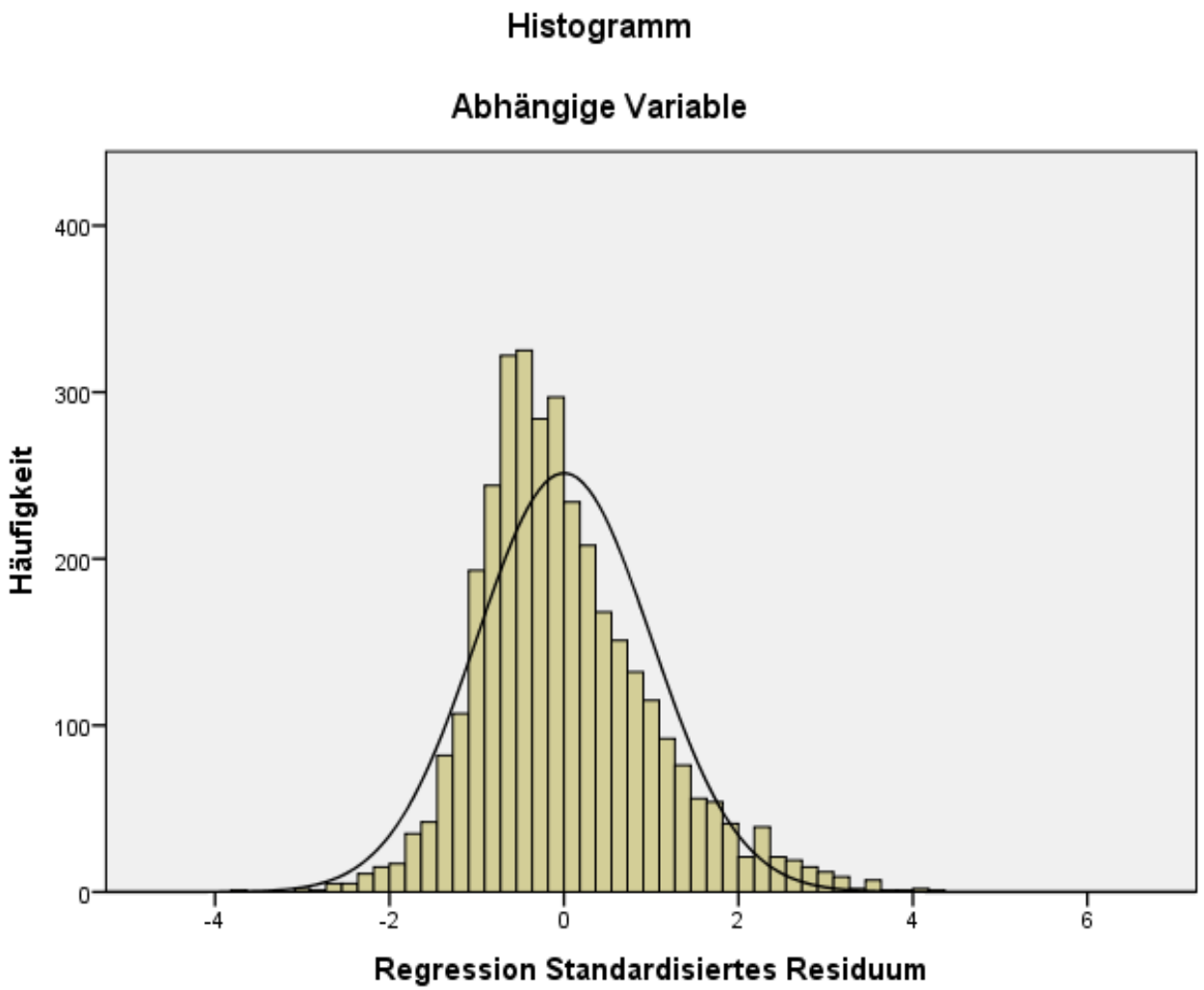
   1. Residuen als Variable Speichern: Im Fenster der linearen Regression auf Speichern -> Residuen (standard. oder unstandard.)

   2. Menu Analysieren -> Deskriptive Statistiken -> explorative Datenanalyse -> Variable mit den gespeicherten Residuen ins Feld der AV ziehen -> Diagramme: Normalverteilungsdiagramm mit Tests, evtl. zusätzlich das Histogramm

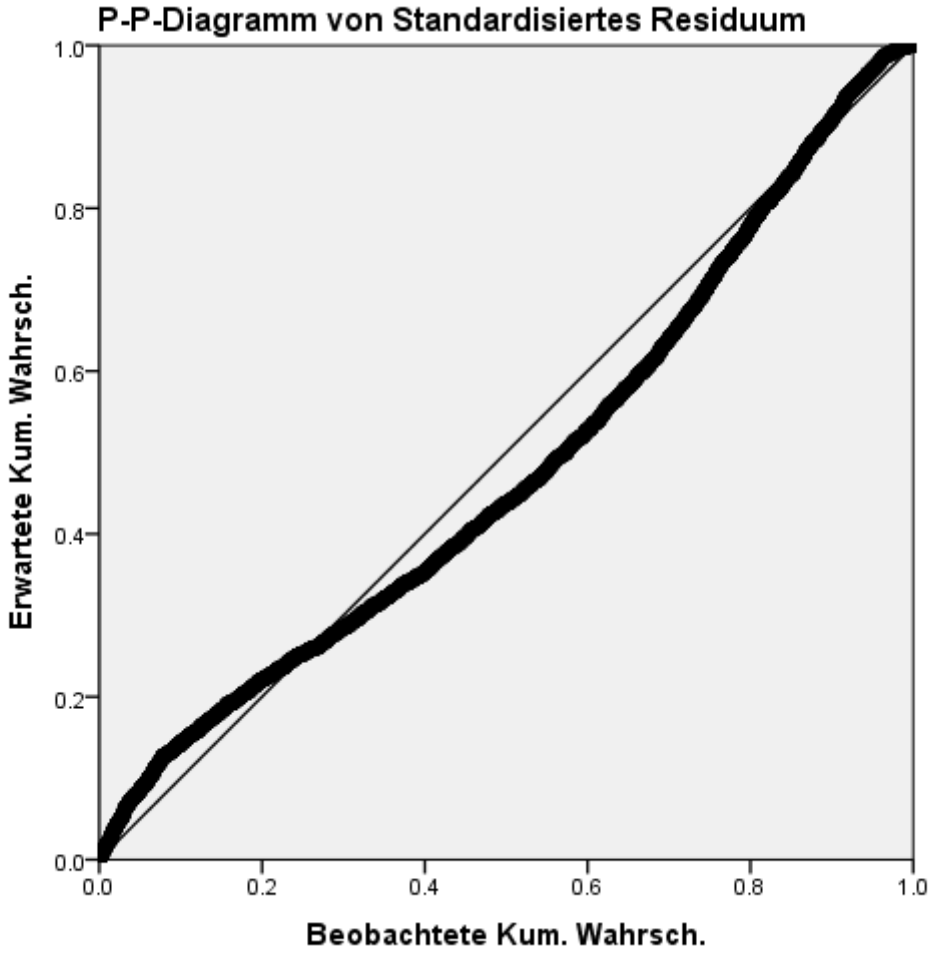
Histogramm Rohwerte



Histogramm Residuen



Normalverteilungsdiagramm



Tests auf Normalverteilung

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
ZRES Standardized Residual, AV: Beschw, UV: PsychProb, Alter18, Sport_01	.074	3465	.000	.967	3465	.000

a. Signifikanzkorrektur nach Lilliefors

- Problem: Werden schnell signifikant bei grossen Stichproben, aber nicht bei kleinen Stichproben (zu wenig Teststärke)

- Aber: Abweichungen von der Normalverteilung sind gerade bei grossen Stichproben (ab n=40) meist nicht problematisch, bei kleinen jedoch schon (Backhaus, 2008, S. 90; Field, 2013, S. 184)*

*Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2008). *Multivariate Analysemethoden* (12. Aufl.). Berlin: Springer.
Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4. Aufl.). Los Angeles: Sage.

Nichtlinearität

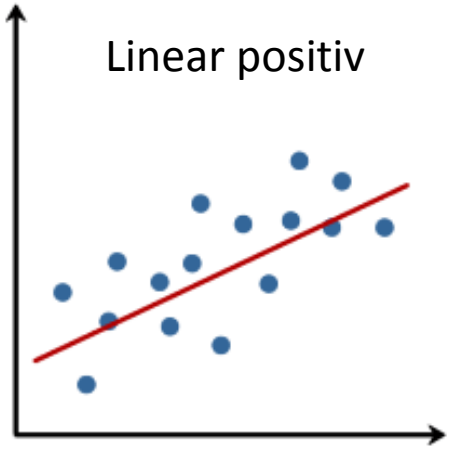
-> Folge: Verzerrung der Schätzwerte

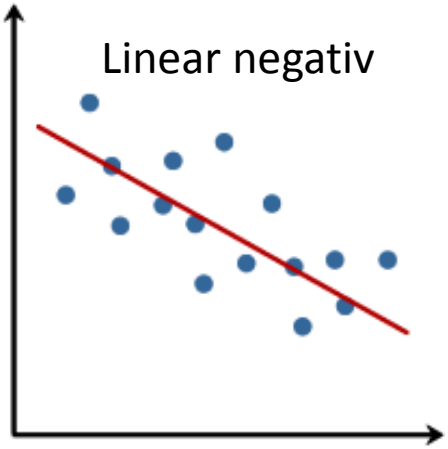
- Kein linearer Zusammenhang zwischen UV und AV
- Visuelle Inspektion Streudiagramm (Scatterplot)

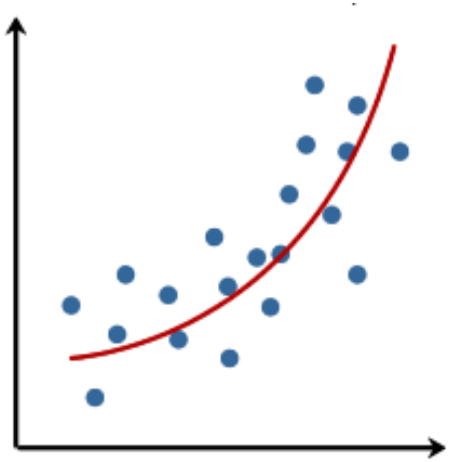
Menu Grafik -> Streu-/Punktediagramm -> UV und AV in die Felder ziehen

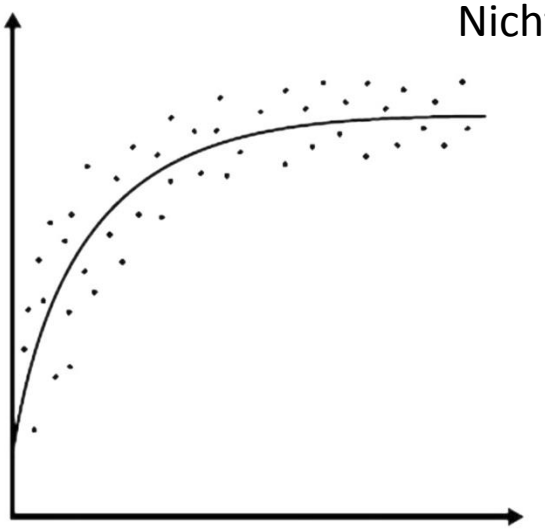
- Nur bivariat möglich (also immer nur die AV und eine UV)
- Linearität wird meist aus der Theorie abgeleitet und bei den Analysen vorausgesetzt

Streudiagramm





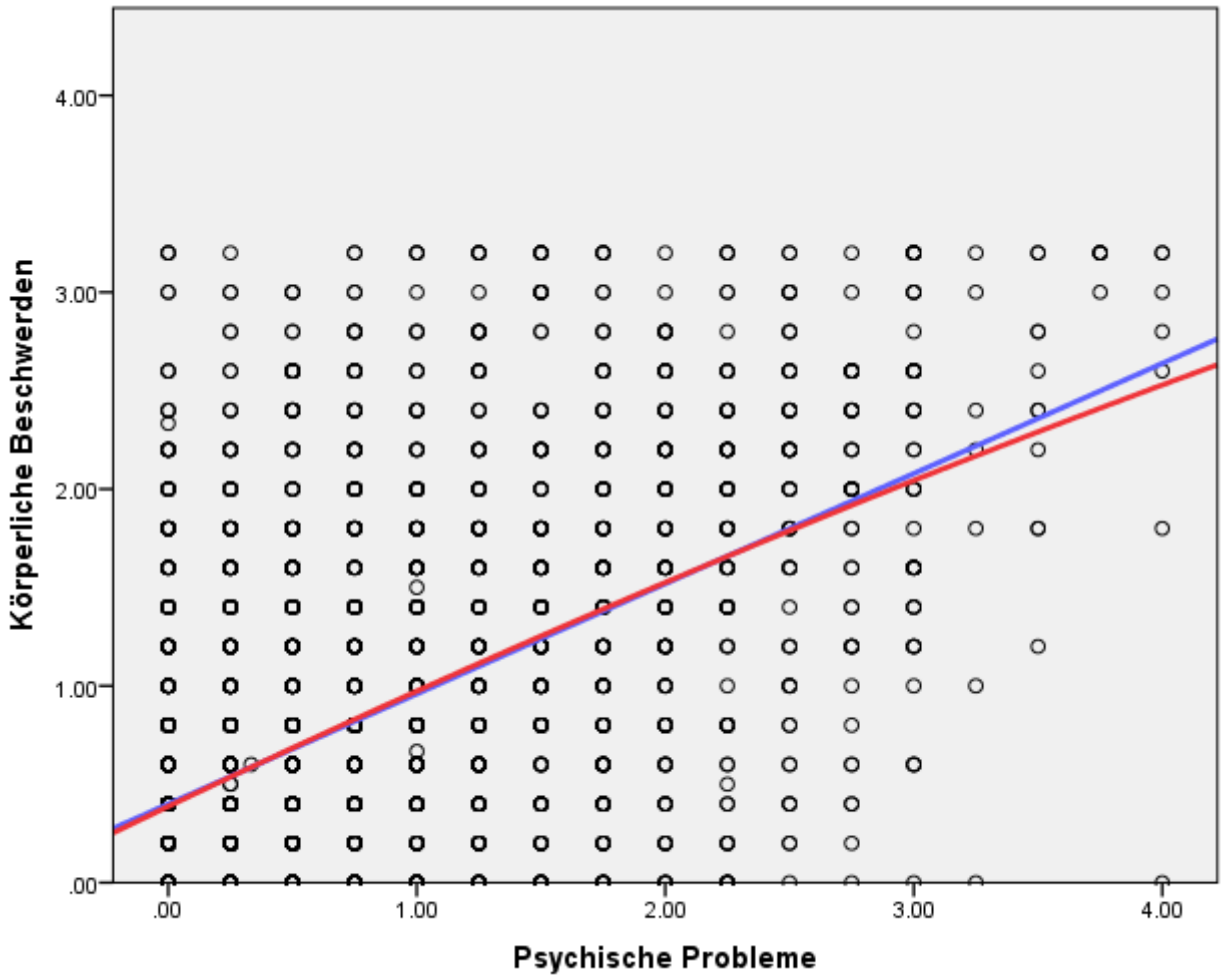




Streudiagramm

Grafik -> Diagrammerstellung -> Streu-/Punktdiagramm -> metrische Variablen auf die X- resp. die Y-Achse ziehen -> ok
-> Ausgabe: Doppelklick auf Grafik zum Bearbeiten -> Anpassungslinie bei Gesamtsumme hinzufügen

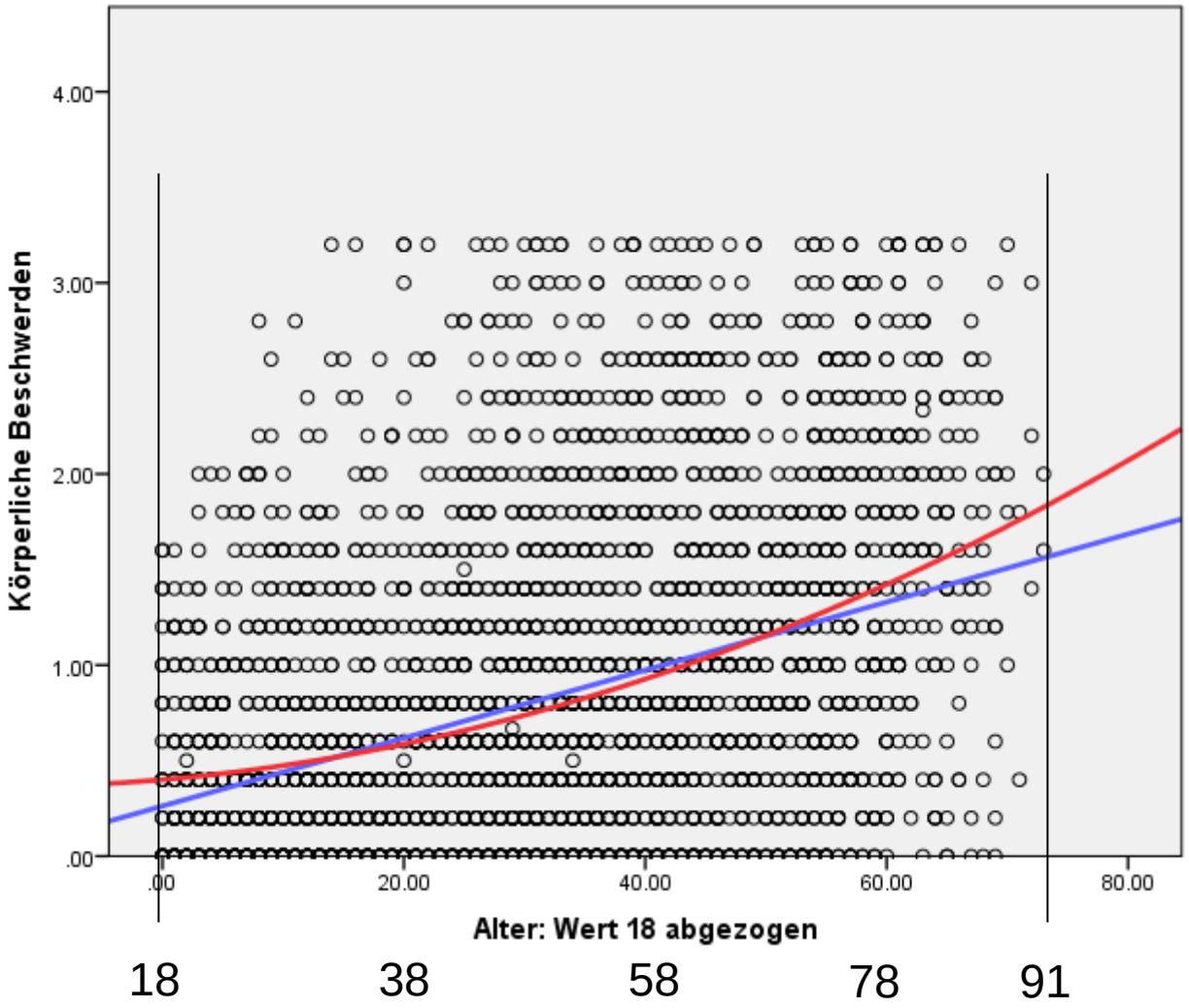
-> linear
-> quadratisch



Streudiagramm

Grafik -> Diagrammerstellung -> Streu-/Punktdiagramm -> metrische Variablen auf die X- resp. die Y-Achse ziehen -> ok
-> Ausgabe: Doppelklick auf Grafik zum Bearbeiten -> Anpassungslinie bei Gesamtsumme hinzufügen

-> linear
-> quadratisch



Heteroskedastizität (Inhomogenität der Varianzen)

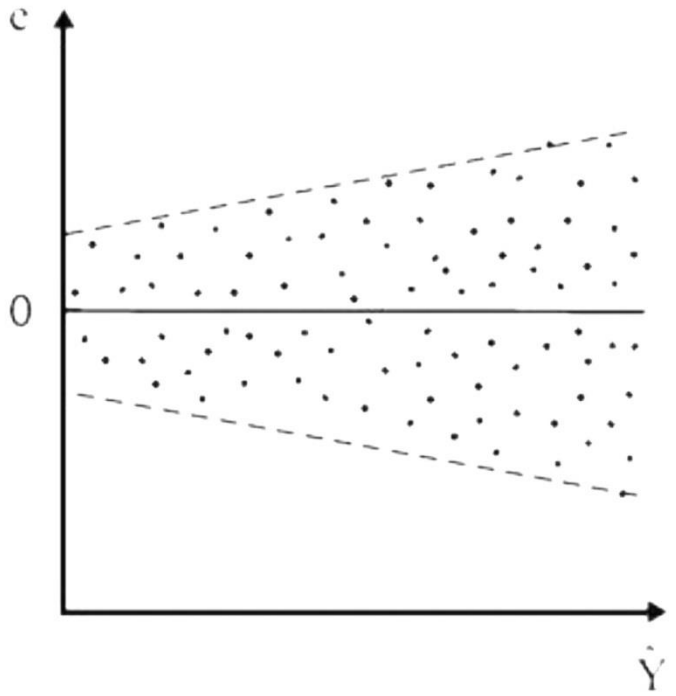
-> Folge: Ungültige Signifikanztests

- Wenn die Residuen nicht homogen Streuen über alle Ausprägungen der vorhergesagten Werte hinweg (Gesamtmodell) -> Varianz ist bei verschiedenen Ausprägungen unterschiedlich gross
- Dasselbe gilt, wenn anstelle des Gesamtmodells der Zusammenhang zwischen der AV und den einzelnen UVs (unter Kontrolle der anderen UVs) betrachtet wird (Partialkorrelation)
- Meist ersichtlich anhand einer Trichterform im Streudiagramm

Im Fenster der linearen Regression auf Diagramme -> Y = ZRESID, X = ZPRED; alle partiellen Diagramme erzeugen

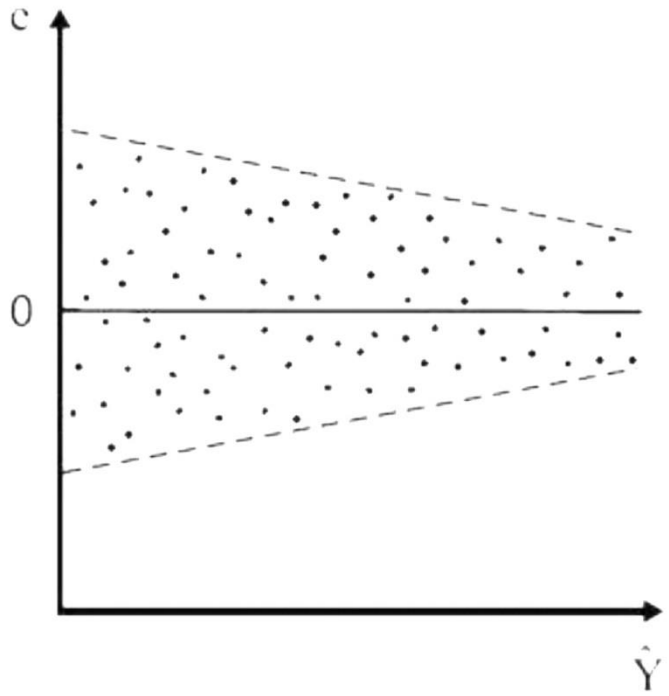
Streudiagramm zwischen vorhergesagten Werten und Residuen: Beispiel Heteroskedastizität

Res



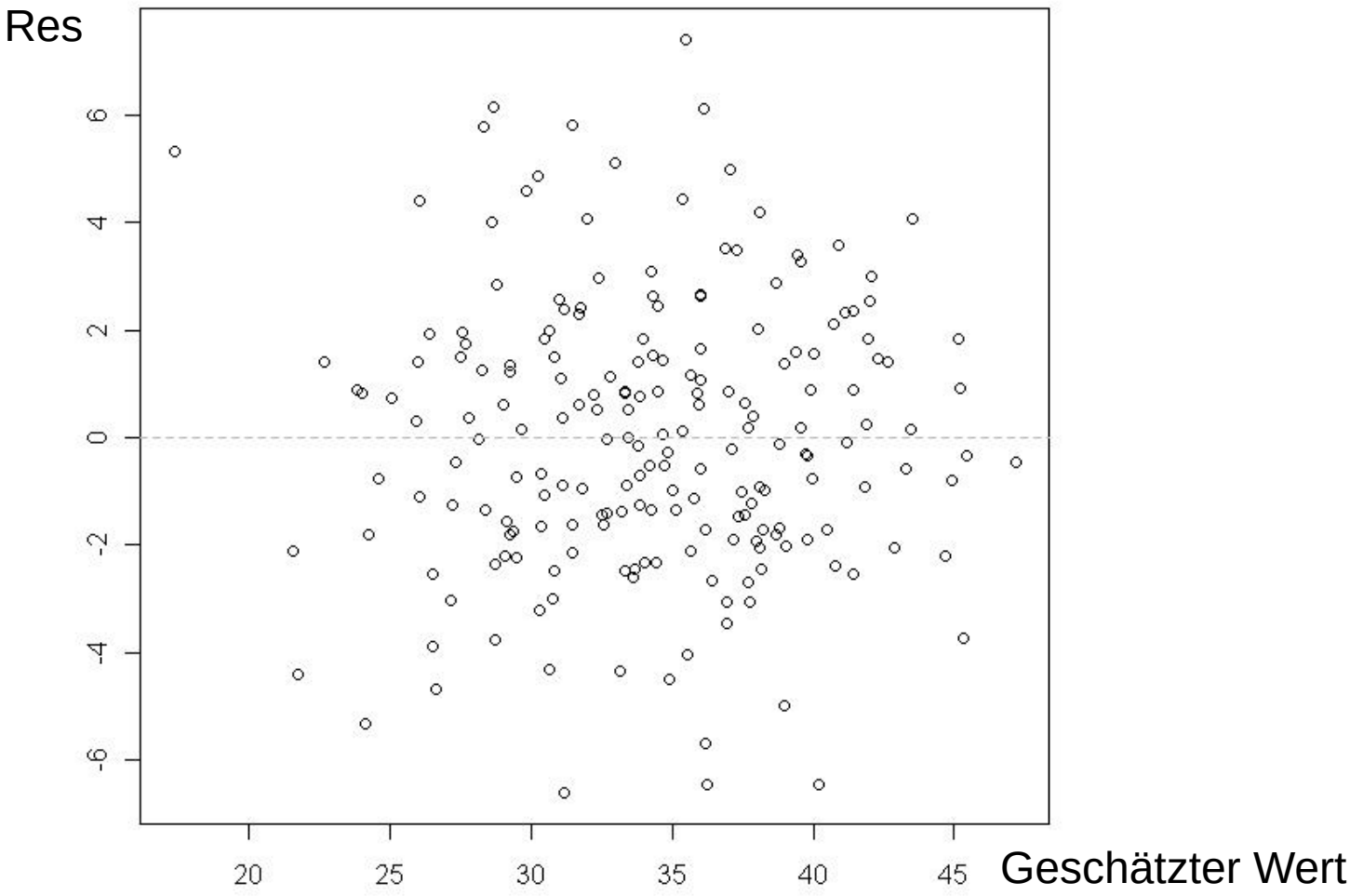
Geschätzter Wert

Res

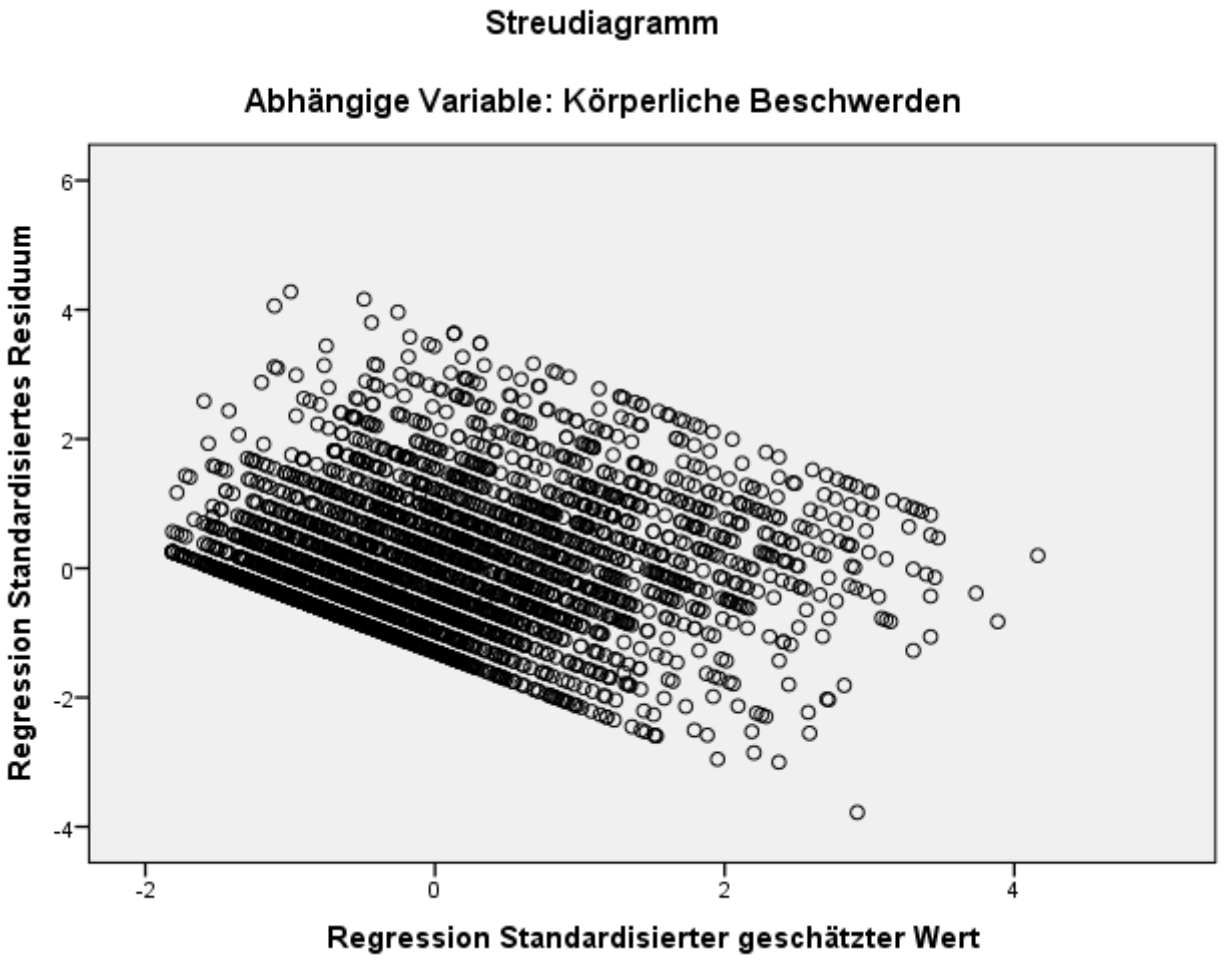


Geschätzter Wert

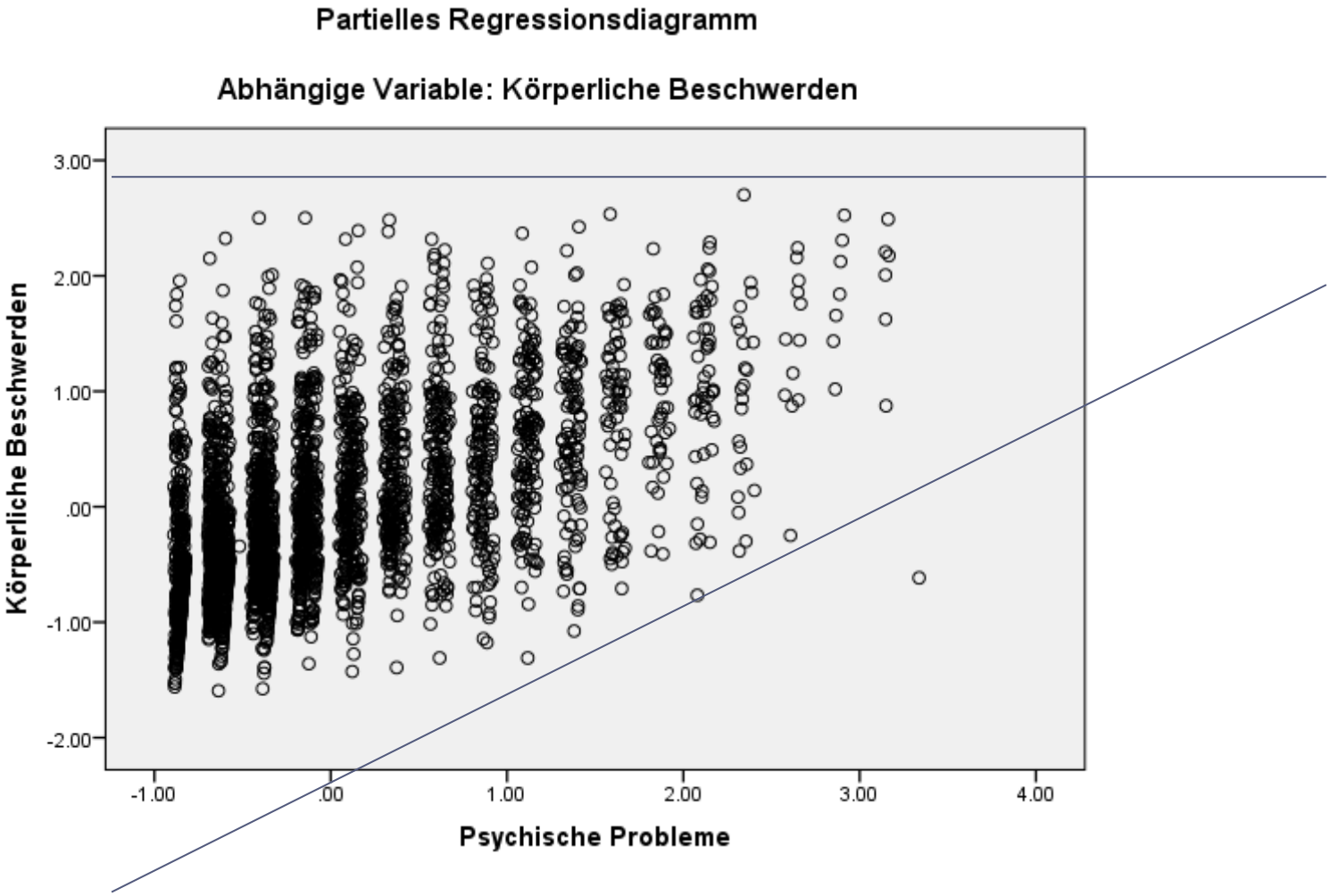
Streudiagramm zwischen vorhergesagten Werten und Residuen: Beispiel Homoskedastizität



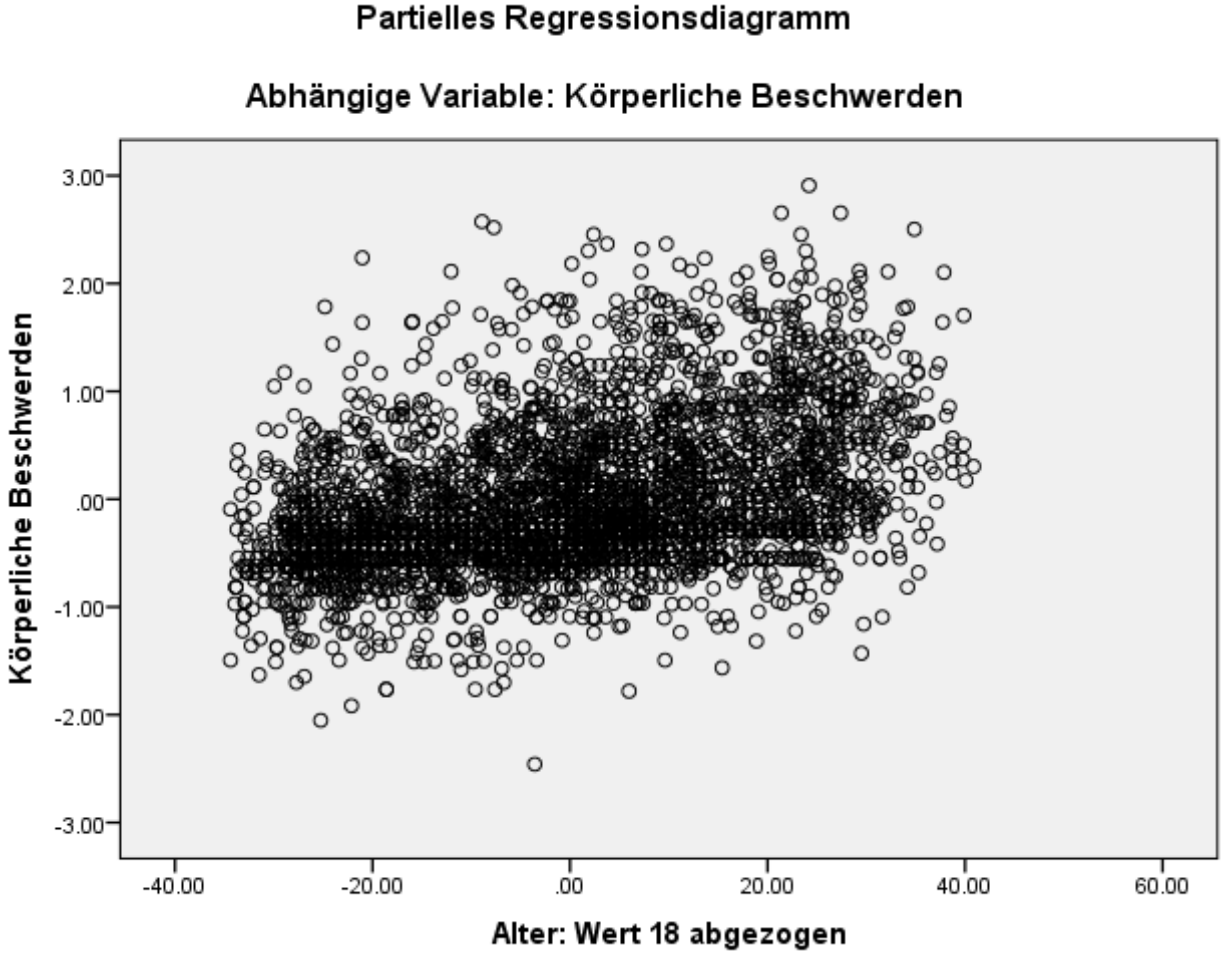
Streudiagramm zwischen vorhergesagten Werten und Residuen bzgl. körperlicher Beschwerden



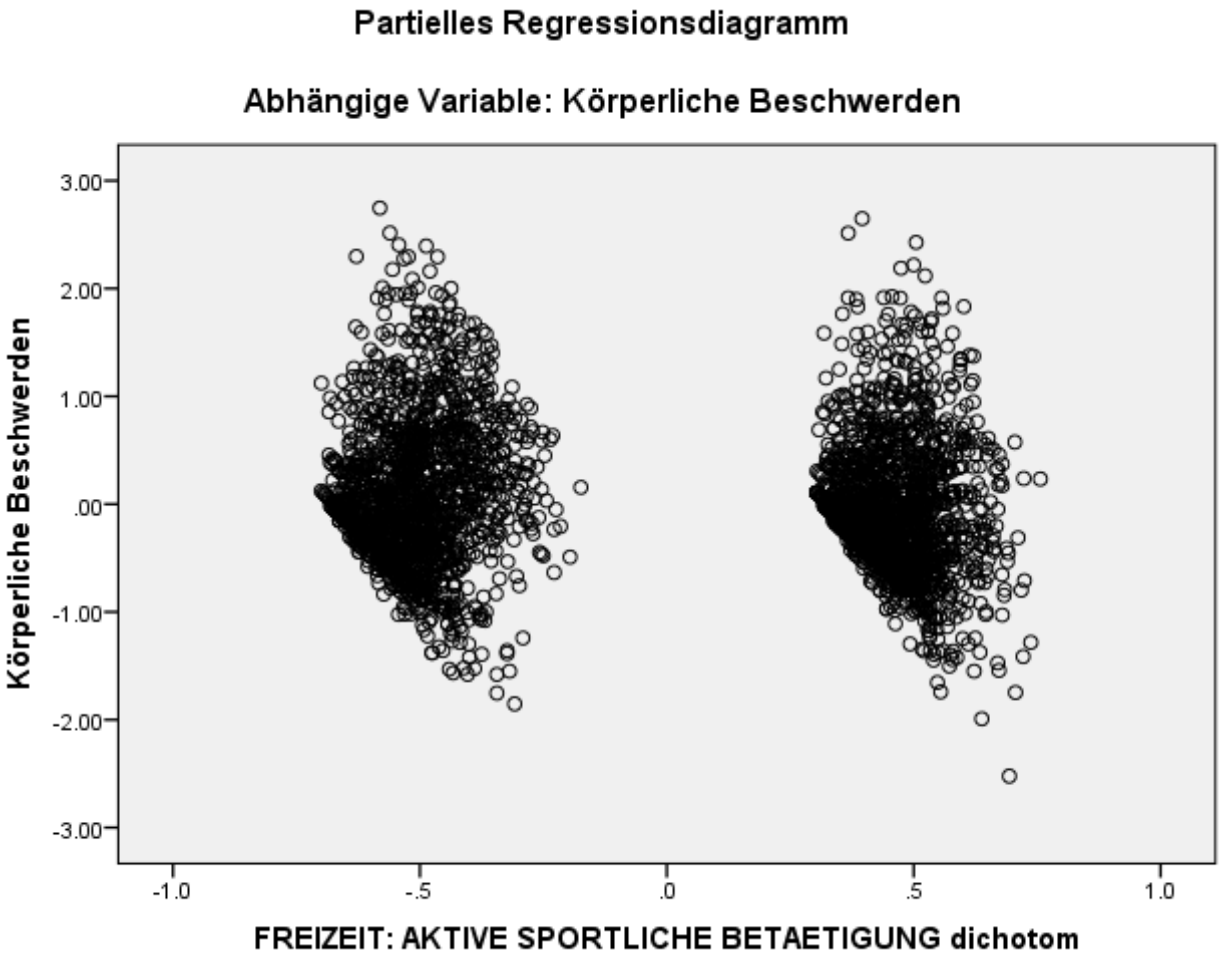
Streudiagramm zwischen vorhergesagten Werten und Residuen bzgl. körperlicher Beschwerden



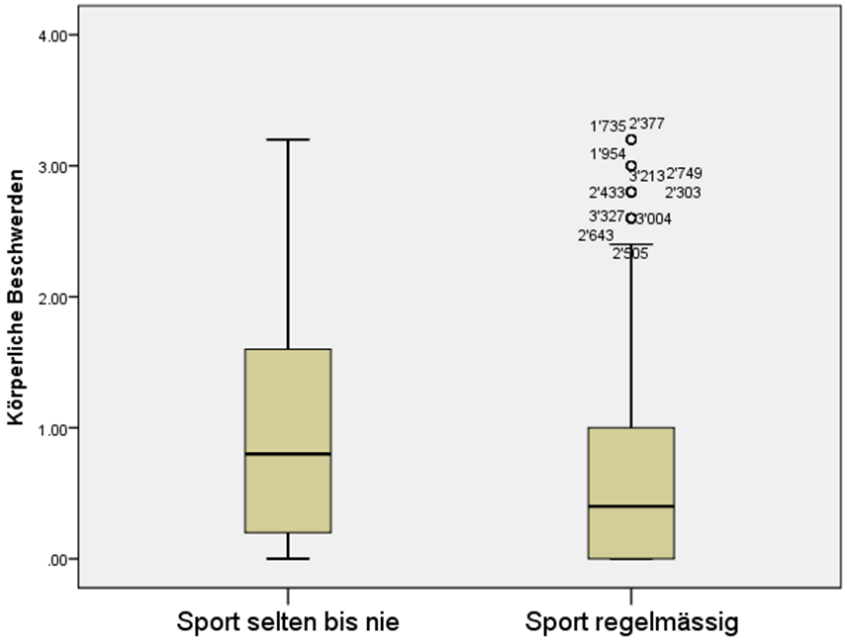
Streudiagramm zwischen vorhergesagten Werten und Residuen bzgl. körperlicher Beschwerden



Streudiagramm zwischen vorhergesagten Werten und Residuen bzgl. körperlicher Beschwerden



Homogenität der Varianzen bei dichotomen Variablen



	Sport_01 FREIZEIT: AKTIVE SPORTLICHE BETAETIGUNG dichotom		Statistik
Beschw Körperliche Beschwerden	0 Sport selten bis nie	Mittelwert	1.0096
		Median	.8000
		Varianz	.836
		Standardabweichung	.91417
	1 Sport regelmässig	Mittelwert	.6517
		Median	.4000
		Varianz	.510
		Standardabweichung	.71388

Homogenität der Varianzen bei dichotomen Variablen

Test auf Homogenität der Varianz

		Levene-Statistik	df1	df2	Signifikanz
Beschw Körperliche Beschwerden:	Basiert auf dem Mittelwert	161.420	1	3466	.000
	Basiert auf dem Median	132.082	1	3466	.000
	Basierend auf dem Median und mit angepaßten df	132.082	1	3455.231	.000
	Basiert auf dem getrimmten Mittel	164.088	1	3466	.000

Weicht signifikant von der Homogenität ab

Analysieren -> deskriptive Statistiken -> explorative Datenanalyse -> Diagramme: Streubreite vs. Mittleres Niveau mit Levene –Test, nicht transformiert

Autokorrelation

-> Folge: Ungültige Signifikanztests

- Wenn die Residuen nicht unabhängig sind, sondern korreliert
- Vor allem bei Modellen mit Messwiederholung ein Problem, wenn die Werte der einzelnen Messzeitpunkte nicht unabhängig sind (Korrelation mit sich selbst zu einem früheren Zeitpunkt)
- Auch möglich bei Klusterung von Individuen in Gruppen
- Durbin-Watson-Test: Werte von 0-4, sollte in der Mitte bei 2 liegen

Im Fenster der linearen Regression auf Statistiken -> Durbin-Watson

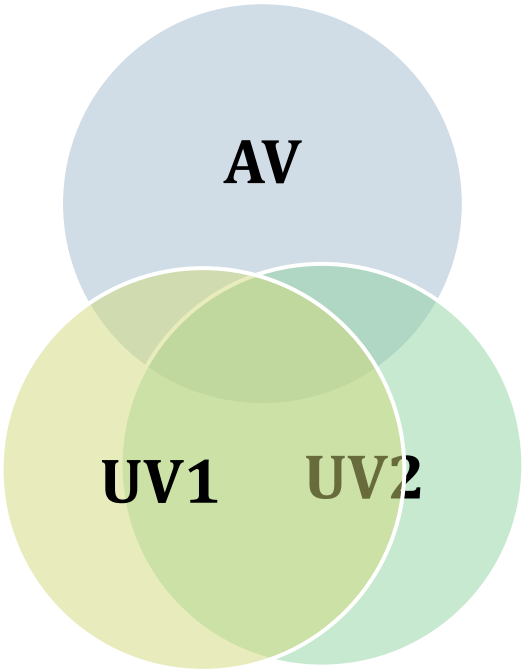
Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers	Durbin-Watson-Statistik
1	.649[a]	.421	.421	.63531	2.032

Multikollinearität

-> Folge: Verminderte Präzision der Schätzwerte

- Wenn die verschiedenen UVs miteinander korreliert sind
- In der multiplen Regression fast unvermeidbar
- Erst problematisch, wenn zu hoch: heben sich gegenseitig auf
- Ersichtlich durch Korrelation zwischen den UVs und den Kennwerten Toleranz und VIF (variance inflation factor)
- Wenn Toleranz < 0.25 und VIF > 5
  -> Hinweis auf problematische Multikollinearität



Multikollinearität

1. Korrelation: Menu Analysieren -> Korrelation -> bivariat -> Pearson
2. Toleranz/VIF: Im Fenster der linearen Regression auf Statistiken -> Kollinearitätsdiagnose

Korrelation / Toleranz / VIF bzgl. körperlicher Beschwerden

Korrelationen

		PsychProb	Sport_01	Alter18
PsychProb	Korrelation nach Pearson	1	-.119**	.036*
	Signifikanz (2-seitig)		.000	.032
	N	3469	3468	3466
Sport_01	Korrelation nach Pearson	-.119**	1	-.140**
	Signifikanz (2-seitig)	.000		.000
	N	3468	3470	3467
Alter18	Korrelation nach Pearson	.036*	-.140**	1
	Signifikanz (2-seitig)	.032	.000	
	N	3466	3467	3468

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

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Multikollinearität wenn:
Toleranz < 0.25; VIF > 5

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten			Kollinearitätsstatistik	
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.	Toleranz	VIF
1	(Konstante)	.001	.029		.043	.966		
	PsychProb	.535	.014	.509	39.043	.000	.985	1.015
	Alter18	.016	.001	.340	26.011	.000	.980	1.020
	Sport_01	-.177	.022	-.106	-8.054	.000	.967	1.034

a. Abhängige Variable: Beschw Körperliche Beschwerden: MEAN(V227_2,V228_2,V233_2,V235_2,V236_2)

Verletzung Modellprämissen: Was nun?

Prämissenverletzung	Lösungsstrategie
Nicht-normalverteilte Residuen	- Ab N = 40 meist kein Problem (fraglich bei extremen Abweichungen, v.a. bei Ausreissern) - **Bootstrapping** - *Transformation der AV (z.B. Logarithmieren) resp. Modellierung einer speziellen Verteilung (-> SPSS: verallgemeinerte lineare Modelle)*
Nichtlinearität	- *Transformation der UV, z.B. Quadrierung -> $Y = a + bx^2$* - Berücksichtigung von Interaktionseffekten (Wechselwirkungen) zwischen UVs
Heteroskedastizität	- Kann mit Nichtlinearität zusammenhängen (-> Massnahmen Nichtlin.) - **Bootstrapping**
Autokorrelation der Residuen	- *Berücksichtigung der Clusterung (Zeitpunkte in Individuen, Individuen in Gruppen etc. -> Mehrebenenanalyse)* - **Bootstrapping**
Multikollinearität	- Entfernen einer UV, die mit einer anderen hoch korreliert ist - Zusammenfassen zu einem Index

Bootstrapping

- Schätzverfahren, welches keine Prämissen erfordert; Alternative zu nicht-parametrischen Tests
- Annahme parametrischer Tests: Die statistischen Kennwerte, welche auf Signifikanz geprüft werden, sind normalverteilt
- Ist diese Annahme nicht haltbar, wissen wir die Verteilung dieser Kennwerte nicht -> Signifikanztests sind ungültig
- Mittels Bootstrapping wird die Verteilung der Kennwerte aus den vorhandenen Daten geschätzt (muss jetzt keine Normalverteilung mehr sein)

Bootstrapping

- Die beobachteten Daten dienen als Population, aus welcher immer wieder Stichproben gezogen werden
- Die Zahl der Ziehungen kann definiert werden (können mehrere Tausend sein)
- Auswirkung auf Ergebnisse: 2 Beispiele

  (1) Datensatz ALLBUS: Vorhersage körperlicher Beschwerden durch psychische Probleme, Alter und sportliche Aktivität

  (2) Fiktiver Datensatz mit N=30: Vorhersage der Häufigkeit ärztlicher Konsultationen durch das Alter

Koeffizienten[a]

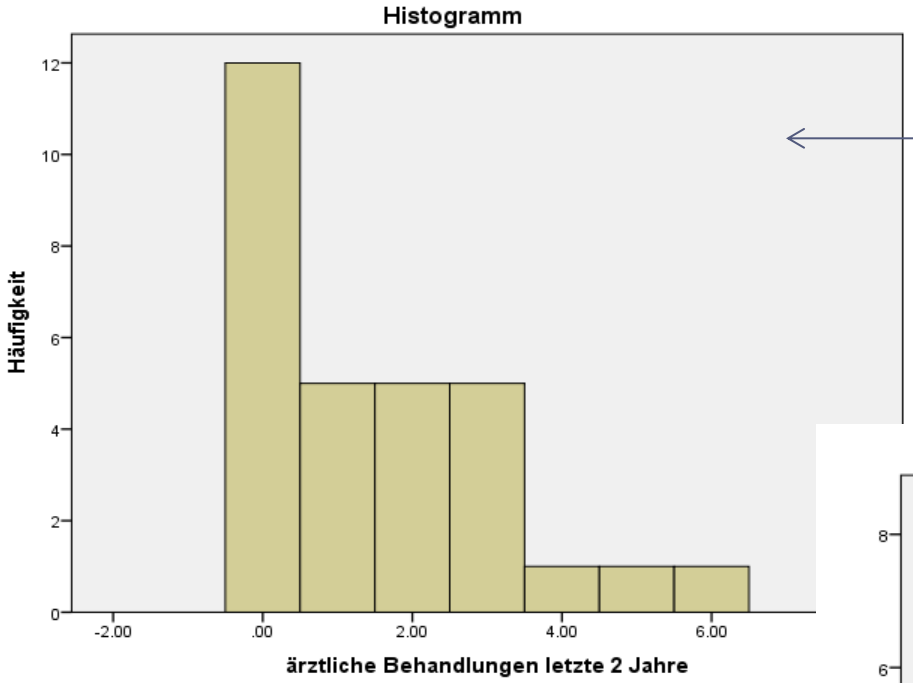
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	.001	.029		.043	.966
	PsychProb	.535	.014	.509	39.043	.000
	Alter18 Alter: Wert 18 abgezogen	.016	.001	.340	26.011	.000
	Sport_01	-.177	.022	-.106	-8.054	.000

a. Abhängige Variable: Beschw Körperliche Beschwerden

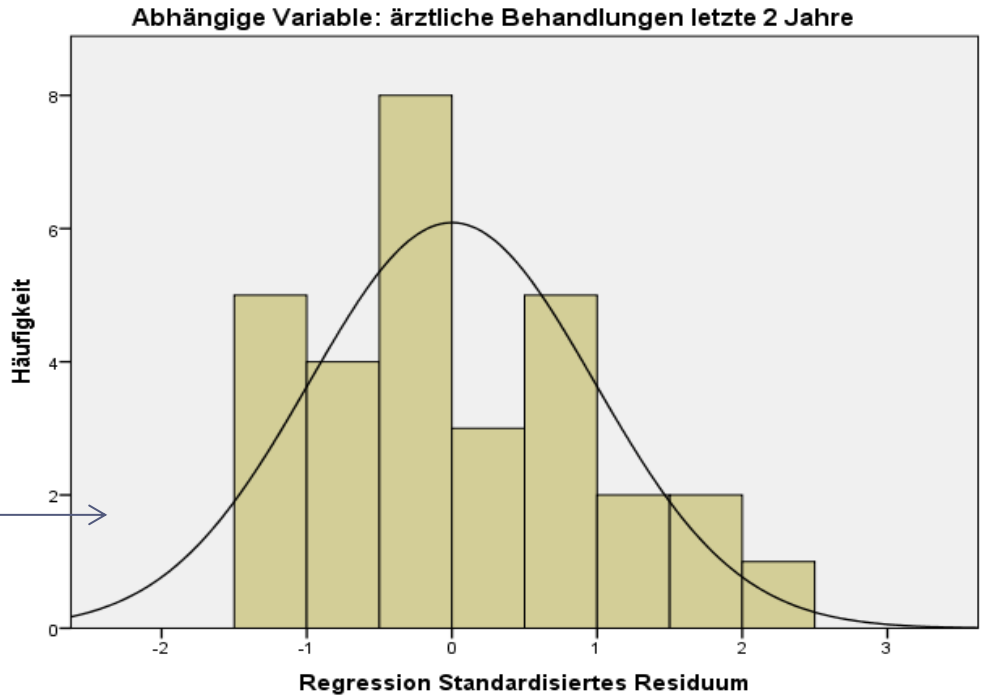
Bootstrap für Koeffizienten

Modell		Regressionskoeffizient B	Bootstrap[a]				
			Verzerrung	Standardfehler	Sig. (2-seitig)	95% Konfidenzintervall	
						Unterer Wert	Oberer Wert
1	(Konstante)	.001	.000	.025	.955	-.045	.050
	PsychProb	.535	.000	.015	.001	.506	.563
	Alter18 Alter: Wert 18 abgezogen	.016	-2.479E-5	.001	.001	.015	.017
	Sport_01	-.177	.001	.023	.001	-.219	-.131

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben



Histogramm Rohwerte der AV ärztliche Behandlungen



Histogramm Residuen der AV ärztliche Behandlungen, vorhergesagt durch das Alter

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		Regressionskoeffizient B	Standardfehler	Beta	T	Sig.
1	(Konstante)	.015	.725		.020	.984
	Alter	.031	.014	.388	2.227	.034

a. Abhängige Variable: Ärztl_Beh ärztliche Behandlungen letzte 2 Jahre

Bootstrap für Koeffizienten

Modell		Regressionskoeffizient B	Bootstrap[a]				
			Verzerrung	Standardfehler	Sig. (2-seitig)	95% Konfidenzintervall	
						Unterer Wert	Oberer Wert
1	(Konstante)	.015	-.011	.770	.991	-1.530	1.438
	Alter	.031	-.001	.016	.065	-.002	.060

a. Sofern nicht anders angegeben, beruhen die Bootstrap-Ergebnisse auf 1000 Bootstrap-Stichproben

Ausblick: Logistische Regression

- AV = kategorial (nominal / ordinal); UVs = versch. Skalenniveaus möglich
- Die Frage ist: Welche Faktoren (UVs) erhöhen die Wahrscheinlichkeit, dass eine bestimmte Ausprägung der AV eintritt
- Bspw. die Aussonderung von Schüler(inne)n mit Verhaltensschwierigkeiten: