der Bundeswehr Universität München

Fakultät für Humanwissenschaften
Sozialwissenschaftliche Methodenlehre
Prof. Dr. Daniel Lois

Lineare Regression: Grundlagen und BLUE-Annahmen

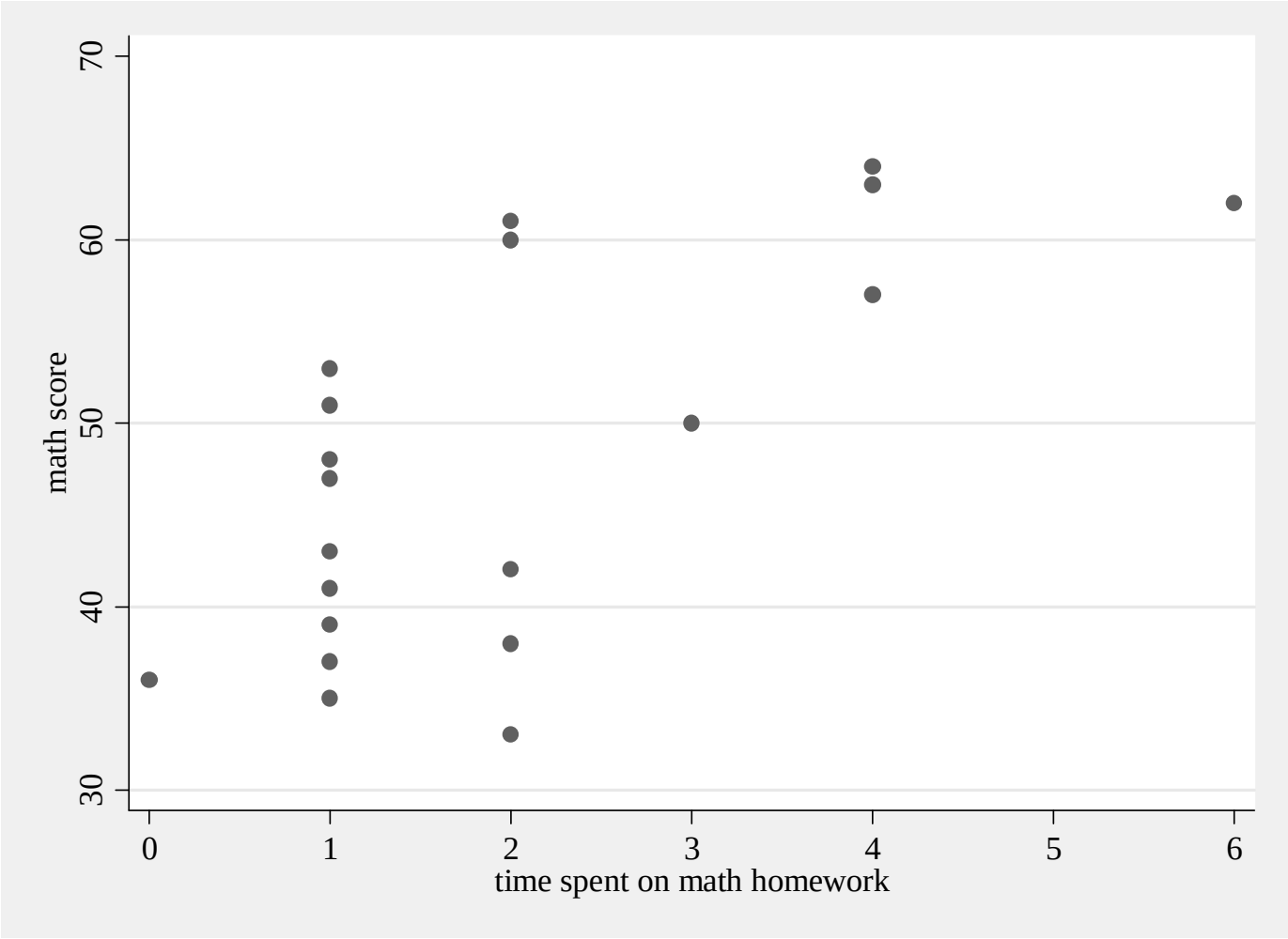
Stand: Juni 2015 (V2.0)

Inhaltsverzeichnis

Lineare Regression: Grundlagen

- **Lineare Regression**: Verfahren zur Analyse des Einflusses von einer oder mehreren unabhängigen Variablen, die ein beliebiges Messniveau aufweisen können, auf eine metrische abhängige Variable
- Das Prinzip wird anhand eines Beispiels verdeutlicht, das in dem folgenden Streudiagramm dargestellt ist
- Auf der y-Achse ist das Ergebnis eines standardisierten Leistungstest für 8.-Klässler im Fach Mathematik dargestellt (abhängige Variable) und auf der x-Achse die Zeit in Wochenstunden, die ein Schüler für Mathe-Hausaufgaben aufwendet (unabhängige Variable)
- Die Lage der Punktewolke deutet darauf hin, dass es sich um einen positiven Zusammenhang handelt: Je mehr Zeit für Hausaufgaben, desto besser das Testergebnis

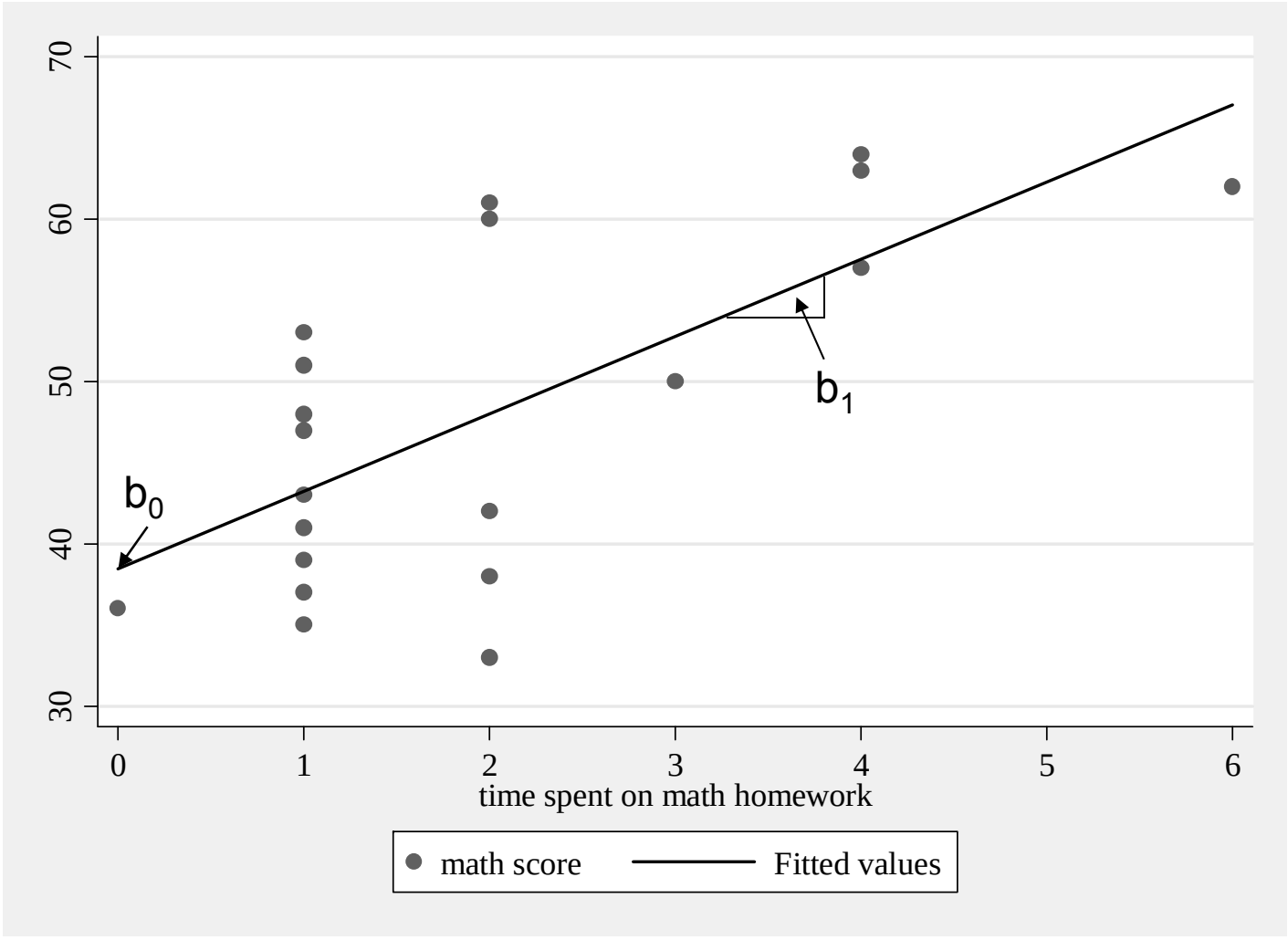
Lineare Regression: Grundlagen



Lineare Regression: Grundlagen

- Die lineare Regression ist ein asymmetrisches Verfahren, d.h. eine Variable wird als abhängig definiert und durch eine oder mehrere unabhängige Variable erklärt
- Die lineare Regression basiert darauf, die beobachteten Werte, die im Streudiagramm dargestellt wurden, möglichst gut durch ein statistisches Modell (eine Gerade) abzubilden
- In die Punktewolke der Beobachtungswerte wird also eine Gerade eingezeichnet, auf dieser Geraden liegen die Vorhersagewerte

Lineare Regression: Grundlagen



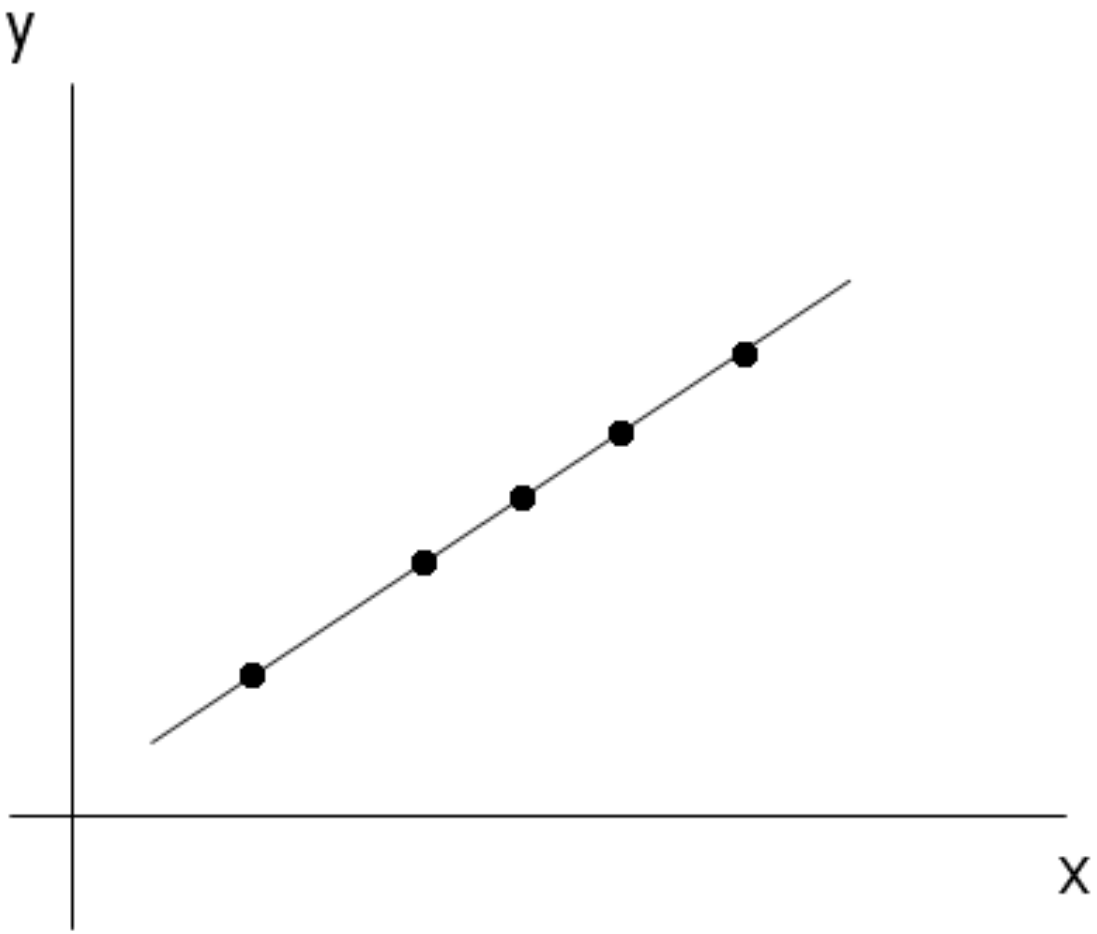
Lineare Regression: Grundlagen

- Die (hier noch unvollständige) Gleichung der bivariaten linearen Regression, durch welche Position und die Steigung der Geraden festgelegt werden, lautet:

$$y_i = b_0 + b_1 x_i$$

- y ist die vorherzusagende (abhängige) Variable für Schüler (i = 1,2,…n), b_0 die Regressionskonstante (auch: Achsenabschnitt, „intercept“), b_1 das Regressionsgewicht („slope“) und x_i eine unabhängige Variable
- Wie wird nun die Position der Linie in der Punktewolke bestimmt?
- Wenn alle Punkte auf einer Geraden liegen würden, dann wäre dies die „best mögliche“ Gerade, da sie alle Punkte repräsentiert; bei der Vorhersage von y durch x würden also keine Fehler gemacht (siehe nächste Folie)

Lineare Regression: Grundlagen



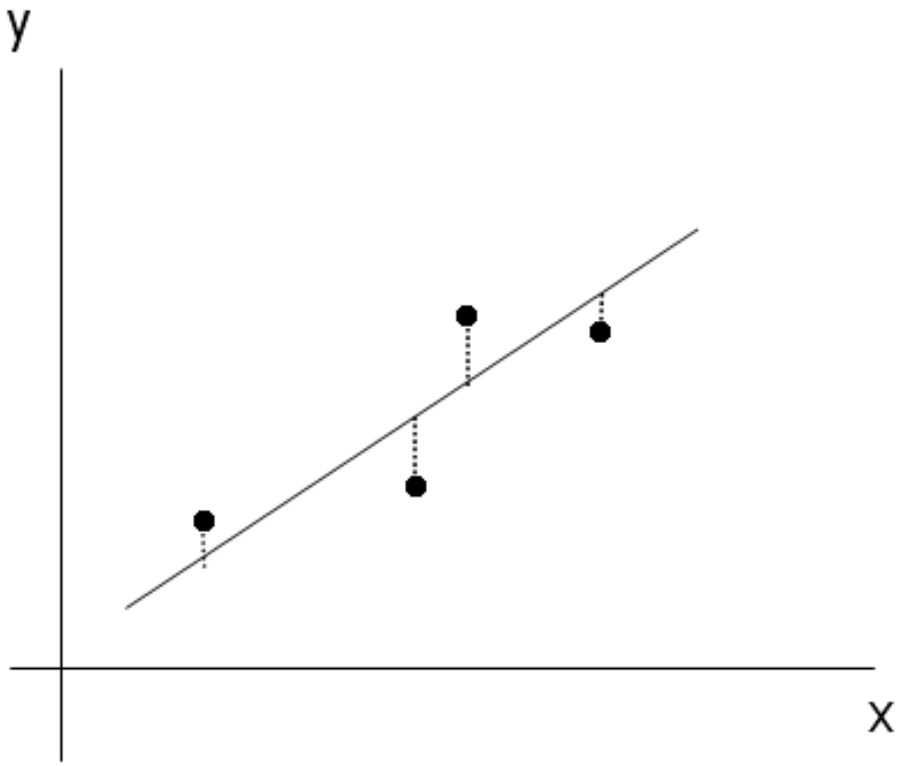
Lineare Regression: Grundlagen

- In der Praxis werden jedoch bei der Vorhersage von y durch x praktisch immer Fehler gemacht; die vollständige bivariate Regressionsgleichung lautet daher:

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

- e_i ist ein Fehlerterm, der durch die Abweichung zwischen Vorhersage- und Beobachtungswerten (sog. Residuen) geschätzt wird
- Wie wird nun die Gerade an die Punktewolke angepasst? Am besten angepasst könnte z.B. bedeuten, dass die Summe der positiven und negativen Differenzen zwischen Vorhersage- und Beobachtungswerten minimiert wird
- Diese Summe ist jedoch immer null, da sich positive und negative Abweichungen zwischen Vorhersage- und Beobachtungswerten ausgleichen

Lineare Regression: Grundlagen



$$\sum e_i = 0$$

Lineare Regression: Grundlagen

- Minimiert werden daher die quadrierten Abweichungen zwischen Beobachtungs- und Vorhersagewerten ($\hat{y}_i$):

$$\sum_{i=1}^{n} e_i^2 = \min \quad = \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min$$

- Diese Vorgehensweise wird Methode der kleinsten Quadrate bzw. OLS-Methode („ordinary least squares“) genannt
- Es werden also diejenigen Werte von b_o und b_1 gesucht, bei denen die folgende Gleichung ein Minimum hat:

$$\min = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

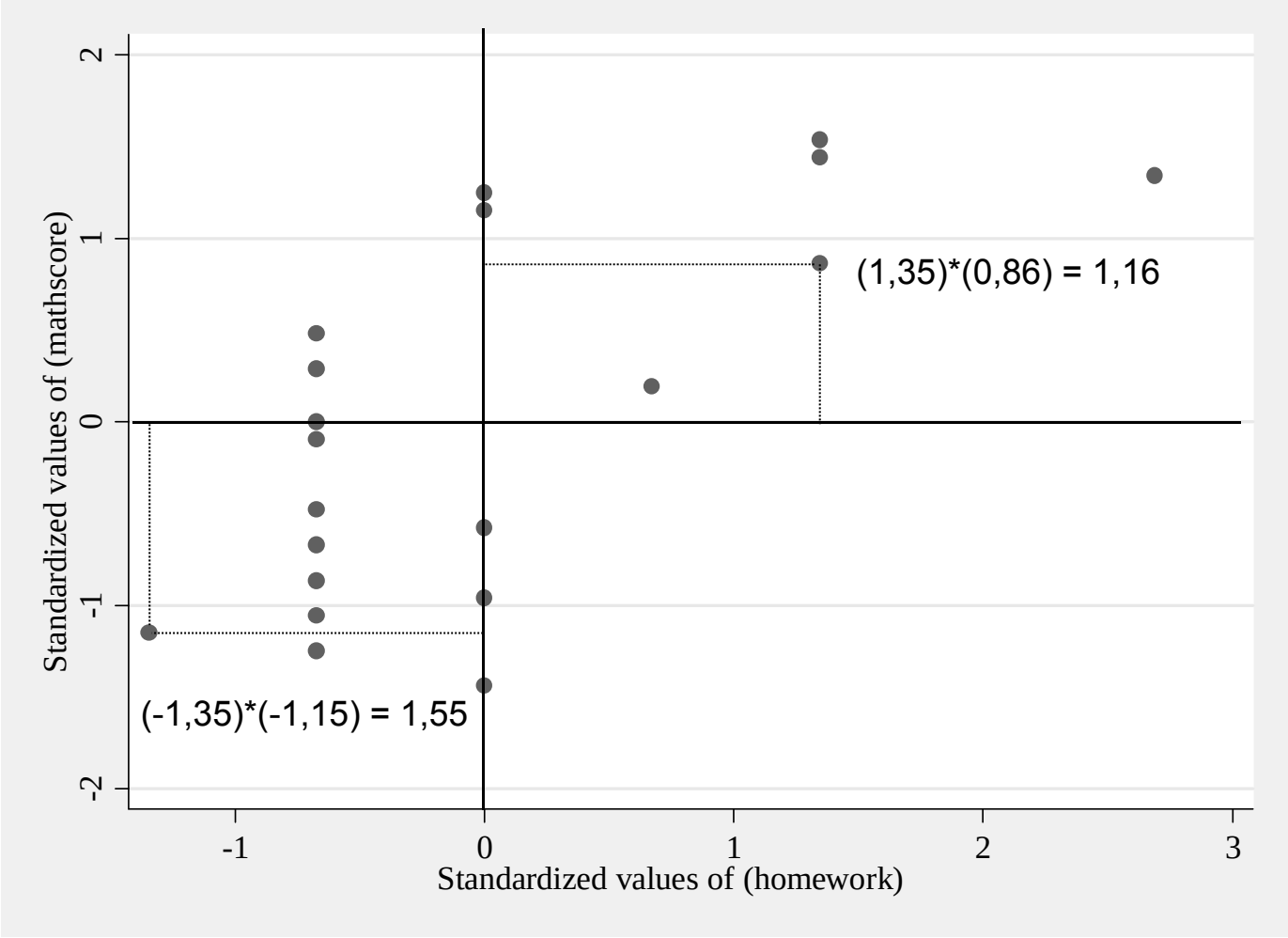
Lineare Regression: Grundlagen

- Wird diese Gleichung nach b_0 und b_1 abgeleitet, folgt daraus:

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Die Formel zeigt, dass b_1 als Quotient der Kovariation der Variablen x und y und der Variation von x berechnet wird
- Die Kovariation ist für das Verständnis der linearen Regression (und → Korrelation, s.u.) zentral und wird daher auf der nächsten Folie grafisch veranschaulicht
- Dort sind die Variablen x und y in z-standardisierter Form dargestellt; d.h., beide Variablen haben einen Mittelwert von 0 und eine Standard-abweichung von 1

Lineare Regression: Grundlagen



Lineare Regression: Grundlagen

- Das Streudiagramm ist anhand der Mittelwerte von x und y in vier Quadranten eingeteilt worden
- Die Kovariation basiert auf dem Produkt der Abweichungen der x- und y-Werte von ihrem jeweiligen Mittelwert: $(x_i - \bar{x})(y_i - \bar{y})$
- Alle Punkte, die im oberen rechten oder unteren linken Quadranten liegen, tragen positive Werte zur Kovariation bei
- Beispiel im Diagramm: Schüler mit y = 0,86 und x = 1,35; die Kovariation beträgt hier: (1,35 - 0)*(0,86 - 0) = 1,16
- Alle Punkte, die im unteren rechten oder oberen linken Quadranten liegen, tragen negative Werte zur Kovariation bei (siehe Beispiel unten links)

Lineare Regression: Grundlagen

- Für die Ausprägung des Regressionskoeffizienten b_1 ist nun entscheidend, wie sich die Punkte im Streudiagramm verteilen
- Liegen die meisten Beobachtungswerte in den Quadranten oben links oder unten rechts, wäre die Kovariation der Variablen y und x in der Summe aller Beobachtungswerte negativ
- In diesem Fall besteht zwischen y und x ein negativer Zusammenhang, was durch einen negativen Regressionskoeffizienten b_1 zum Ausdruck kommt
- Liegen die Beobachtungswerte dagegen, wie im Beispiel, überwiegend in den Quadranten unten links bzw. oben rechts, ist die Kovariation in der Summe positiv und auch b_1 nimmt einen positiven Wert an (je mehr Zeit für Hausaufgaben, desto besser die Leistung)

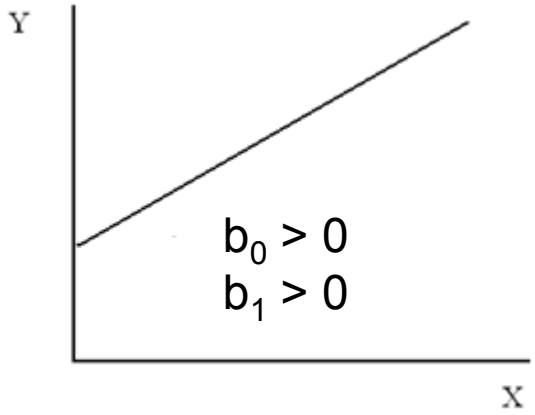
Lineare Regression: Grundlagen

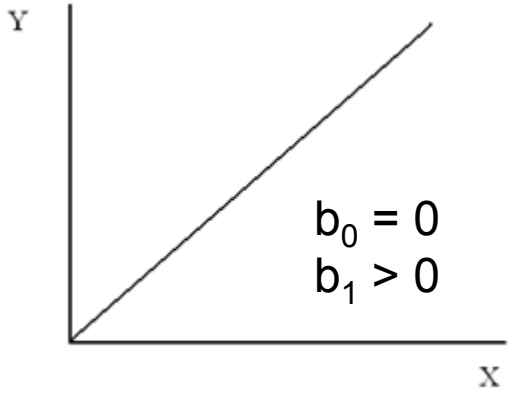
- Ist $b_1 = 0$, ist auch die Kovariation von x und y null und es besteht kein linearer Zusammenhang zwischen den Variablen; die Regressionsgerade verläuft dann parallel zur x-Achse, hat also keine Steigung
- Exakt ist der Regressionskoeffizient b_1 so zu interpretieren, dass sich die Vorhersagewerte des Regressionsmodells für y genau um b_1 Einheiten erhöhen, wenn sich die unabhängige Variable x um eine Einheit erhöht
- b_1 wird auch als unstandardisierter Regressionskoeffizient bezeichnet
- Er gibt in jedem Fall die Richtung des Effekts von x auf y an, sagt jedoch nur bedingt etwas über die Effektstärke aus (→ Beta, s.u.)

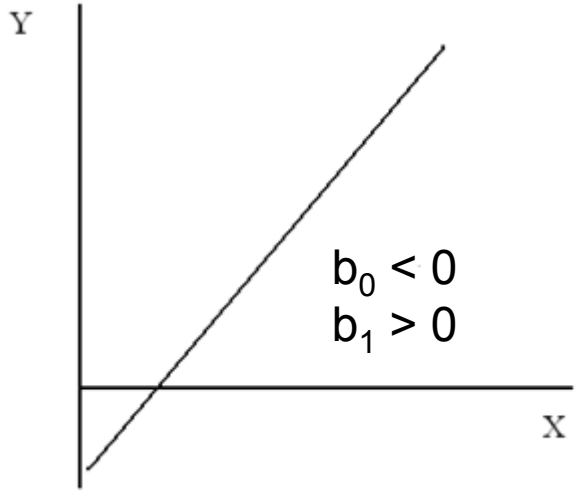
Lineare Regression: Grundlagen

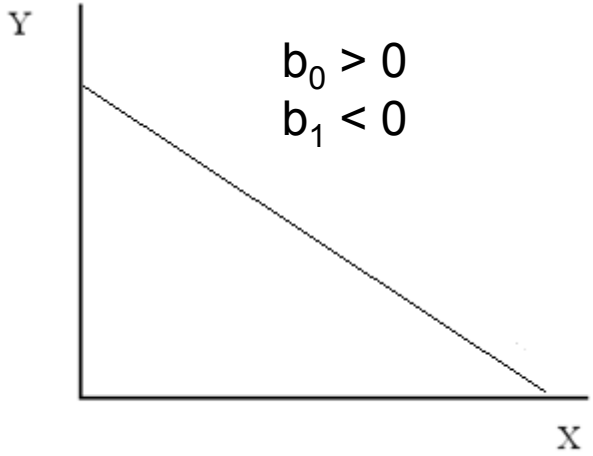
- Die Regressionskonstante b_0 gibt den Schnittpunkt der Regressionsgeraden auf der y-Achse beim Wert x = 0 an (Achsen-abschnitt)
- Bei $b_0 = 0$ schneidet die Gerade die vertikale y-Achse beim Wert x = 0 (sie geht „durch den Ursprung“)
- Ob die Regressionskonstante inhaltlich sinnvoll interpretierbar ist, hängt davon ab, ob der Wert x = 0 zum gültigen Wertebereich gehört
- Im Beispiel ist dies der Fall; x = 0 bedeutet hier, dass der jeweilige Schüler keine Mathematikhausaufgaben macht
- Die nächste Folie verdeutlicht die Lage der Regressionsgeraden bei unterschiedlichen Werten von b_0 und b_1

Lineare Regression: Grundlagen









Lineare Regression: Grundlagen

- Zur Berechnung der Regressionsparameter b_0 und b_1 wird die Arbeitstabelle auf der folgenden Folie benötigt
- Die x-Variable entspricht der Zeit für Hausaufgaben und y entspricht der abhängigen Variablen Mathematikleistung
- Weiterhin dargestellt werden für jede Person die Abweichungen von x und y von ihren jeweiligen Mittelwerten, die Variation von x (vierte Spalte von links) und die Kovariation von x und y
- Die Kovariation ist entscheidend für die Richtung des Regressions-koeffizienten b_1
- Im Beispiel ist die Kovariation in der Summe positiv, b_1 ist somit ebenso positiv, mit steigender Hausaufgabenzeit erhöht sich die Leistung

Lineare Regression: Grundlagen

y_i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
33	2	0	0	-15	0
35	1	-1	1	-13	13
36	0	-2	4	-12	24
37	1	-1	1	-11	11
38	2	0	0	-10	0
39	1	-1	1	-9	9
41	1	-1	1	-7	7
42	2	0	0	-6	0
43	1	-1	1	-5	5
47	1	-1	1	-1	1
48	1	-1	1	0	0
50	3	1	1	2	2
51	1	-1	1	3	-3
53	1	-1	1	5	-5
57	4	2	4	9	18
60	2	0	0	12	0
61	2	0	0	13	0
62	6	4	16	14	56
63	4	2	4	15	30
64	4	2	4	16	32
			$\sum 42$		$\sum 200$

$\bar{y} = 48{,}0 \quad \bar{x} = 2{,}0$

Lineare Regression: Grundlagen

- Nun können die Regressionsparameter ausgerechnet werden:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{200}{42} = 4{,}762$$

$$b_0 = \bar{y} - b_1\bar{x} = 48 - (4{,}762 * 2) = 38{,}476$$

Lineare Regression: Grundlagen

- Die vollständige Regressionsgleichung für diese Regression lautet:

$$y_i = 38{,}476 + (4{,}762 * \text{homework}_i) + e_i$$

- Dies bedeutet:
  - Die vorhergesagte Leistung beträgt 38,476 Punkte wenn x = 0 ist, d.h. wenn der Schüler keine Hausaufgaben macht
  - Erhöht sich die unabhängige Variable um eine Einheit, d.h. macht ein Schüler eine Stunde mehr Hausaufgaben, erhöht sich die Leistung um 4,762 Einheiten
  - e_i erfasst den „Teil“ in der Mathematikleistung, der nicht durch den linearen Effekt der Hausaufgabenzeit erklärt wird

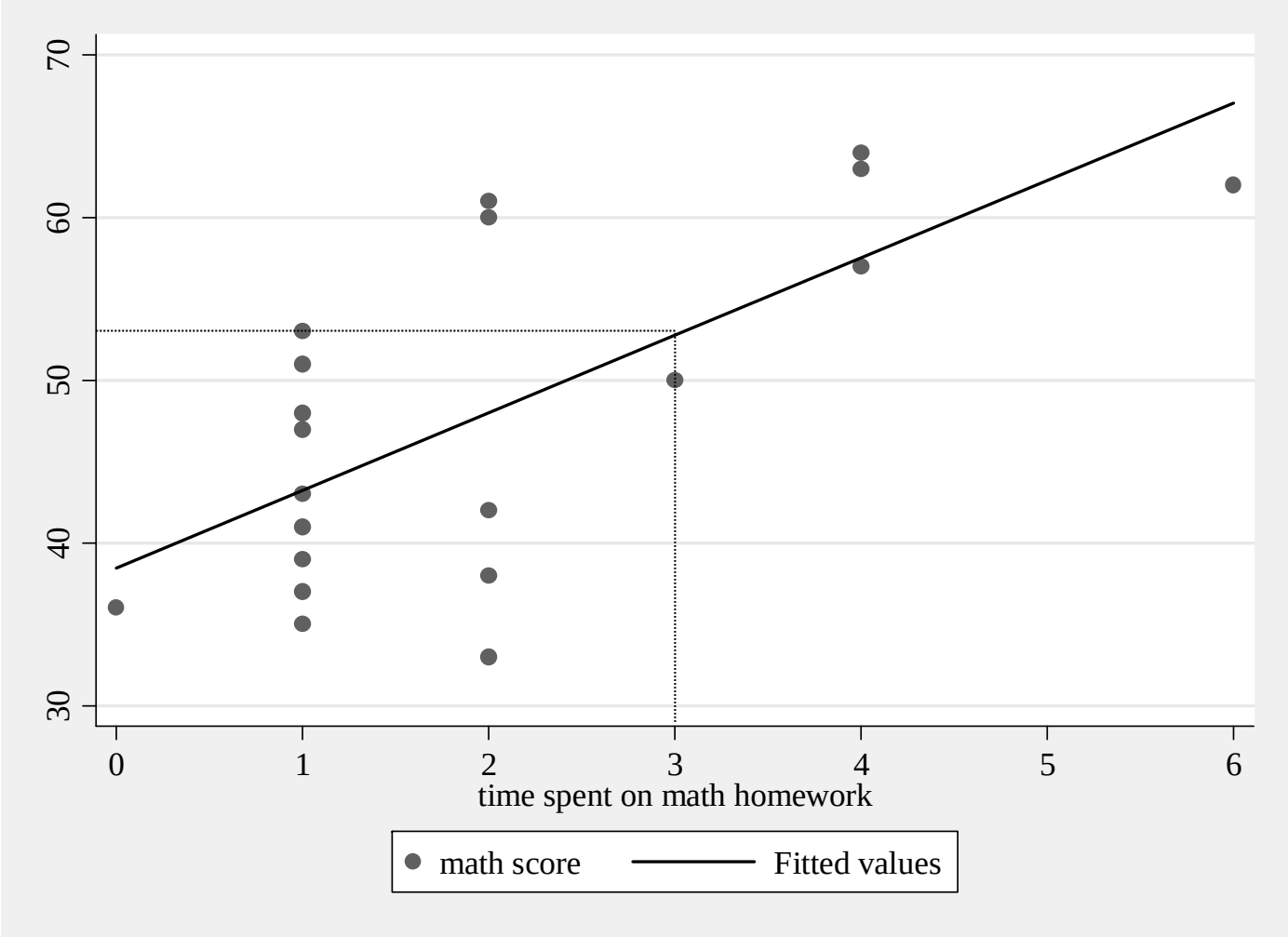
Lineare Regression: Grundlagen

- Über die Regressionsgleichung können nun die y-Vorhersagewerte ausgerechnet werden, die auch die Position der Regressionsgeraden im Koordinatensystem bestimmen
- Zum Beispiel beträgt der y-Vorhersagewert bei x = 3:

$$\hat{y}_{x=3} = 38{,}476 + (4{,}762 * 3) = 38{,}476 + 14{,}29 = 52{,}77$$

- Bei einer Hausaufgabenzeit von 3 Stunden wird durch das Regressionsmodell also eine Leistung von 52,77 vorhergesagt
- Im Koordinatensystem liegt die Regressionsgerade bei einem x-Wert von 3 entsprechend auf dem y-Wert 52,77

Lineare Regression: Grundlagen



Lineare Regression: Grundlagen

- Im Folgenden werden die verschiedenen Kennziffern besprochen, die in SPSS für die bivariate Regression ausgegeben werden
- Als Maß dafür, wie eng die Regressionsgerade an den Punkten der Punktewolke liegt – oder wie gut das Modell an die Daten angepasst ist – wird das Verhältnis zwischen dem erklärten Teil der Streuung und der gesamten Streuung betrachtet (Output ANOVA)
- Bei der nicht erklärten Streuung (in der Gleichung: Fehlerterm bzw. Residuen e_i) handelt es sich um die summierten quadrierten Abweichungen zwischen Vorhersage- und Beobachtungswerten
- Dieser Wert wird unter „Quadratsumme Residuen“ ausgegeben und beträgt hier 1107,6

Lineare Regression: Grundlagen

ANOVA[b]

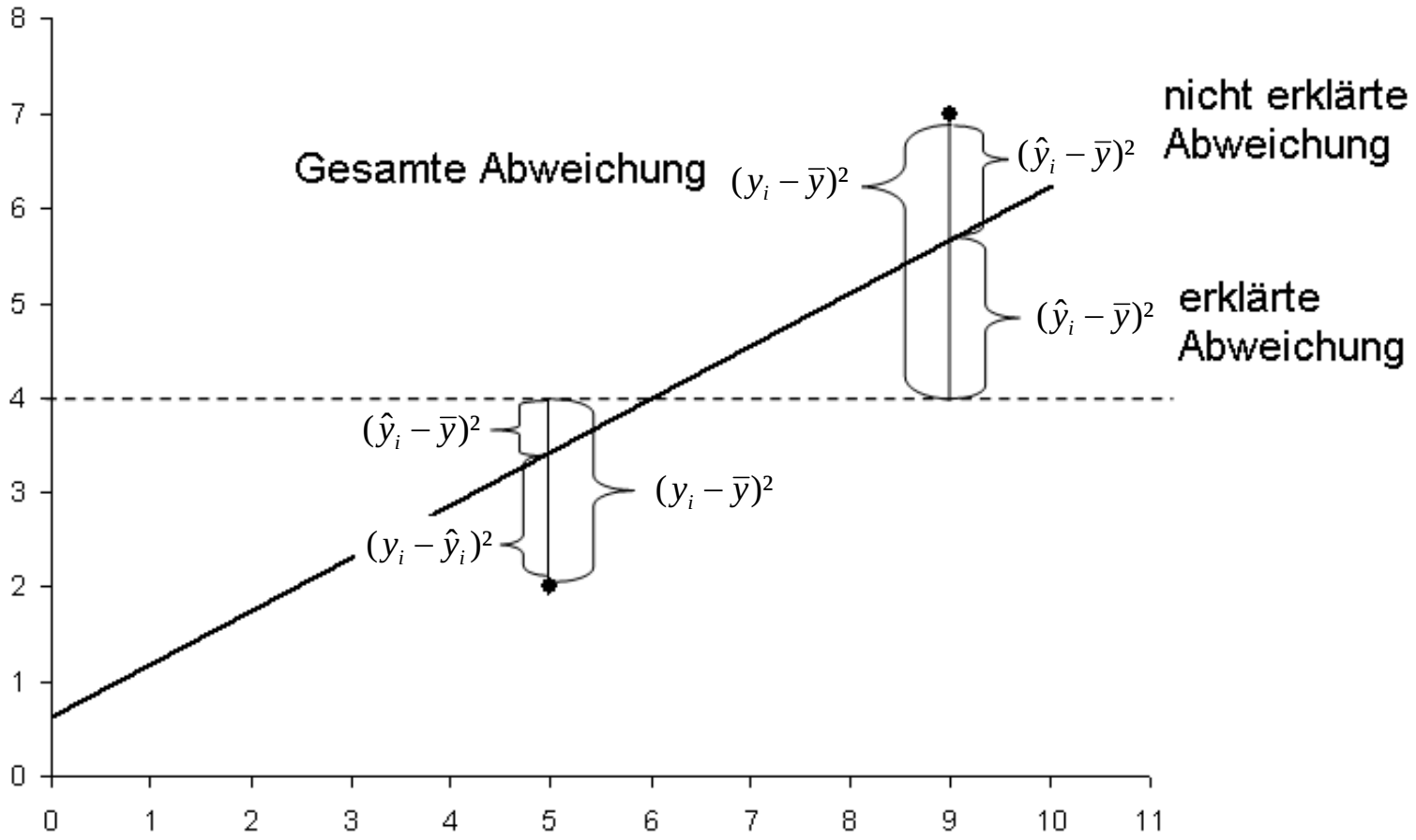
Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	952,381	1	952,381	15,477	,001[a]
	Residuen	1107,619	18	61,534		
	Gesamt	2060,000	19			

a. Einflußvariablen : (Konstante), time spent on math homework

b. Abhängige Variable: math score

- Die erklärte Streuung entspricht den summierten quadrierten Differenzen zwischen Vorhersagewerten und dem Mittelwert von y
- Dieser Wert wird unter „Quadratsumme Regression“ ausgewiesen und beträgt 952,4. Nicht erklärte und erklärte Streuung ergeben zusammen die Gesamtstreuung (2060,0, die summierten quadrierten Abweichungen zwischen y-Mittelwert und den Beobachtungswerten)

Lineare Regression: Grundlagen



Lineare Regression: Grundlagen

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,680[a]	,462	,432	7,84439

a. Einflußvariablen : (Konstante), time spent on math homework

- Das Verhältnis zwischen der Quadratsumme der erklärten Streuung und der Quadratsumme der Gesamtstreuung wird als **R²** (auch: Bestimmtheitsmaß, Fit, Determinationskoeffizient) bezeichnet und ergibt hier:

$$R^2 = \frac{952{,}38}{2060{,}0} = 0{,}462$$

Lineare Regression: Grundlagen

- R^2 folgt einer **PRE („proportional reduction in error“)-Logik**. Alle PRE-Maße basieren auf der Formel: $(E_1 - E_2) / E_1$
- E_1 entspricht der Quadratsumme „Gesamt“ (Fehlersumme, wenn AV durch ihren eigenen Mittelwert vorhergesagt wird)
- E_2 entspricht der Quadratsumme „Residuen“ (Fehlersumme, wenn AV durch die UV (hier: Hausaufgabenzeit) vorhergesagt wird)
- Da (2060 – 1107,6) / 2060 = 0,462, werden bei der Vorhersage der Leistung durch die Hausaufgabenzeit 46,2% weniger Fehler gemacht
- Anders ausgedrückt: 46,2% der Varianz in der Leistung können durch die Hausaufgabenzeit erklärt werden

Lineare Regression: Grundlagen

- Zusätzlich wird ein korrigiertes R^2 ausgegeben, das immer dann zu verwenden ist, wenn das Regressionsmodell mehr als eine unabhängige Variable hat
- Das korrigierte R^2 „bestraft“ komplexe Modelle mit vielen Erklärungsfaktoren und berechnet sich wie folgt (n = Stichprobenumfang, k = Anzahl der Regressionskoeffizienten + Konstante):

$$\text{korr. } R^2 = 1 - \frac{\frac{\text{QS Residuen}}{(n-k)}}{\frac{\text{QS Gesamt}}{(n-1)}}$$

Lineare Regression: Grundlagen

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	952,381	1	952,381	15,477	,001[a]
	Residuen	1107,619	18	61,534		
	Gesamt	2060,000	19			

a. Einflußvariablen : (Konstante), time spent on math homework

b. Abhängige Variable: math score

- Die Quadratsummen werden anhand ihrer Freiheitsgrade (df = degrees of freedom) vergleichbar gemacht
- Die Freiheitsgrade in der Zeile „Regression“ entsprechen der Anzahl der b_1-Koeffizienten
- Die Freiheitsgrade in der Zeile „Residuen“ entsprechen n-2 und in der Zeile „Gesamt“ n-1

Lineare Regression: Grundlagen

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	952,381	1	952,381	15,477	,001[a]
	Residuen	1107,619	18	61,534		
	Gesamt	2060,000	19			

a. Einflußvariablen : (Konstante), time spent on math homework

b. Abhängige Variable: math score

- Der **F-Wert** entspricht dann dem Verhältnis zwischen der erklärten Streuung und der nicht erklärten Streuung
- Die Berechnung lautet: 952,38 / 61,53 = 15,48; die erklärte Streuung ist also 15,5-mal größer als die nicht erklärte Streuung

Lineare Regression: Grundlagen

- Mit Hilfe des F-Wertes wird die Nullhypothese getestet, dass alle Regressionskoeffizienten des Modells in der Grundgesamtheit = 0 sind
- Kann diese Nullhypothese nicht mit hinreichender Sicherheit abgelehnt werden, ist nicht auszuschließen, dass die Regressionskoeffizienten rein zufällig zustande gekommen sind und nicht von der Stichprobe auf die Grundgesamtheit verallgemeinert werden können
- Im Beispiel ist der F-Wert hochsignifikant
- Die Erklärungsleistung des Regressionsmodells ist somit mit hoher Wahrscheinlichkeit nicht rein zufallsbestimmt
- R^2 und der F-Wert sind zusammenfassend Koeffizienten zur Beurteilung des Gesamtmodells

Lineare Regression: Grundlagen

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	38,476	2,990		12,870	,000
	time spent on math homework	4,762	1,210	,680	3,934	,001

a. Abhängige Variable: math score

- Unter B werden zunächst die Konstante (b_0), der nicht standardisierte Regressionskoeffizient (b_1) und dessen Standardfehler ausgegeben
- Es werden genau die Werte angegeben, die weiter oben von Hand berechnet wurden
- Nochmal zur Interpretation von b_0: Wenn der Schüler keine Hausaufgaben macht (x = 0), beträgt die vorhergesagte Leistung 38,476

Lineare Regression: Grundlagen

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	38,476	2,990		12,870	,000
	time spent on math homework	4,762	1,210	,680	3,934	,001

a. Abhängige Variable: math score

- Mit jeder Stunde Mathematik-Hausaufgaben erhöht sich die Leistung um 4,762 (b_1-Koeffizient)
- Da $b_1 > 0$ ist, gibt es einen positiven Effekt der Hausaufgaben auf die Leistung, die Regressionsgerade steigt im Streudiagramm von links unten nach rechts oben

Lineare Regression: Grundlagen

- Eine Schätzung dafür, wie stark verschiedene Regressionskoeffizienten um den wahren Wert streuen (→ Inferenzstatistik), ist der Standardfehler des Regressionskoeffizienten, der wie folgt berechnet wird:

$$s.e.(b_1) = \frac{s_y}{s_x}\sqrt{\frac{1 - r_{xy}^2}{n - 2}}$$

- s_y und s_x sind die Standardabweichungen von x und y, n ist die Stichprobengröße und r^2_{xy} ist die die quadrierte Korrelation zwischen x und y (also das R^2)

Lineare Regression: Grundlagen

- Die Formel zeigt, dass drei Faktoren zu kleinen Standardfehlern beitragen:
  - Ein großer Stichprobenumfang (n)
  - Eine starke Korrelation zwischen x und y
  - Eine hohe Standardabweichung (Streuung) von x

- Für unser Beispiel ergibt sich:

$$s.e.(b_1) = \frac{10,41}{1,49}\sqrt{\frac{1-0,462}{20-2}} = 1,21$$

Lineare Regression: Grundlagen

- Mit Hilfe des Standardfehlers ist es möglich, Rückschlüsse auf die wahre Lage des Regressionskoeffizienten in der Grundgesamtheit zu ziehen (→ Inferenzstatistik, Hypothesentest)
- Der ungünstigste Fall tritt ein, wenn der wahre Regressionskoeffizient b* in der Grundgesamtheit = 0 ist, die unabhängige Variable also tatsächlich keinen Effekt auf die AV hat
- Diese sog. Nullhypothese (der wahre Wert von b_1 ist in der Grundgesamtheit = 0) wird mit Hilfe der **t-Statistik** getestet
- Berechnung: t-Wert durch Standardfehler von b_1

$$t = \frac{b_1}{s.e.(b_1)}$$

Lineare Regression: Grundlagen

- Faustformel: Ab einem Stichprobenumfang von (ca.) n = 100 sind t-Werte ab 2,0 – bzw. (bei negativem b) ab -2,0 – auf dem 95%-Niveau signifikant sind (der exakte p-Wert lässt sich in der Praxis im Output des jeweiligen Statistikprogramms ablesen)
- Dies bedeutet, dass die Nullhypothese (b_1 ist in der Grundgesamtheit = 0) bei t = 2,0 mit einer Sicherheit von etwa 95% und einem Alpha-Fehler-Risiko von 5% abgelehnt werden kann (ab t-Werten von etwa 2,6 beträgt die Sicherheit 99%)

Lineare Regression: Grundlagen

- **Beta** wird im bivariaten Fall wie folgt berechnet (wobei b_1 der Regressionskoeffizient einer unabhängigen Variablen x, s_x die Standardabweichung derselben Variablen und s_y die Standardabweichung der abhängigen Variablen ist):

$$beta = b_1 \frac{s_x}{s_y} = 4{,}762 \frac{1{,}49}{10{,}41} = 0{,}68$$

- Betas rangieren in der Regel (wie → Korrelationen, s.u.) zwischen -1 und 1 und erlauben daher eine eindeutige Beurteilung von Effektstärken – und zwar auch dann, wenn die UVs in unterschiedlicher Metrik gemessen sind
- Da in die Berechnung von Beta Standardabweichungen einfließen, wird diese Kennziffer lediglich für metrische UV empfohlen

Lineare Regression: Grundlagen

- Zusätzlich zum Test des Regressionskoeffizienten gegen 0 kann man sich fragen, in welchem Wertebereich b_1 in der Grundgesamtheit wahrscheinlich liegt (→ Inferenzstatistik, Konfidenzintervalle)
- Den genauen Wert können wir mit Stichprobendaten zwar nicht bestimmen. Es ist jedoch möglich, ein Konfidenzintervall zu schätzen, in dem der wahre Wert mit bestimmter Wahrscheinlichkeit liegt:

$$b_1 \pm \text{t - Wert} * \text{s.e.}(b_1)$$

- Je nach akzeptiertem Alpha-Fehler-Risiko sind als Faustformel t-Werte von 2,0 (~ 95%-Konfidenzintervall) oder 2,6 (~ 99%-Konfidenzintervall) einzusetzen

Lineare Regression: Grundlagen

- Das 95%-Konfidenzintervall für den Hausaufgaben-Effekt im Beispiel beträgt dann:

$$4{,}762 \pm 2{,}0 * 1{,}21$$

- Es ergibt sich das Intervall mit den Grenzen [2,34; 7,18]
- Dies ist eine Stichprobenschätzung für das Intervall, in dem der wahre Effekt von Hausaufgaben auf Leistung mit einer Wahrscheinlichkeit von 95% liegt

Lineare Regression: Grundlagen

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	47,800	2,761		17,315	,000
	Mann	,800	5,521	,034	,145	,886

a. Abhängige Variable: math score

- Übungsbeispiel mit dichotomer UV (1 = Mann, 0 = Frau)
- Die vorhergesagte Leistung der Frauen liegt bei 47,8
- Männer sind demgegenüber um $b_1 = 0{,}8$ besser in Mathematik; der Geschlechtsunterschied ist hier aber nicht signifikant ($t = 0{,}145$, $p = 0{,}886$)

Lineare Regression: Grundlagen

- Bei mehrstufig kategorialen Variablen (z.B. Schultyp mit den 3 Ausprägungen (1.) öffentlich, (2.) privat in religiöser Trägerschaft, (3.) sonstige Privatschulen) wird pro Ausprägung bis auf eine (die Referenzkategorie) eine Dummy-Variable in das Modell aufgenommen
- Interpretation: Mathematikleistung der Schüler in religiösen Privatschulen liegt, gegenüber der Referenzkategorie der öffentlichen Schulen (Durchschnittsleistung 49,8), um $b_1 = 3{,}3$ Einheiten höher
- Schüler in sonstigen Privatschulen durchschnittlich um $b_2 = 10{,}7$ Einheiten besser als Schüler in öffentlichen Schulen
- Ob der Unterschied zwischen den beiden Privatschul-Typen signifikant ist, wird in diesem Modell nicht getestet

Lineare Regression: Grundlagen

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	49,845	,075		661,484	,000
	private_relig	3,298	,187	,116	17,604	,000
	private_other	10,651	,277	,254	38,480	,000

a. Abhängige Variable: math score

BLUE-Annahmen

- Lineare Regressionen liefern nur dann sinnhafte und unverzerrte Ergebnisse, wenn eine Reihe von Voraussetzungen erfüllt sind (sog. **BLUE-Annahmen**, „best linear unbiased estimator“)
- Einige der wichtigsten BLUE-Annahmen werden nun (in SPSS) getestet:
  - Linearität: Die Beziehung zwischen der abhängigen Variablen und der (den) unabhängigen Variablen ist linear
  - Die Residuen folgen bestimmten Regeln: symmetrische Verteilung und Homoskedastizität
  - Es gibt keine Multikollinearität der erklärenden Variablen

BLUE-Annahmen

- Beispieldaten: ALLBUS (Allgemeine Bevölkerungsumfrage in den Sozialwissenschaften, N = 2229 erwerbstätige Personen)
  - Abhängige Variable: Nettoerwerbseinkommen monatlich in EUR
  - Unabhängige Variablen:
    - Geschlecht (Frau = 1, Mann = 0)
    - Berufserfahrung in Jahren
    - Bildungsjahre (8 bis 20)
    - Wohnort: Ostdeutschland (= 1, West = 0)

BLUE-Annahmen

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	-555,129	122,482		-4,532	,000
	Berufserfahrung	7,766	1,313	,114	5,916	,000
	Geschlecht: Frau	-554,229	43,923	-,234	-12,618	,000
	Wohnort: Ostdeutschland	-371,520	45,975	-,149	-8,081	,000
	Bildungsjahre	167,166	7,901	,409	21,158	,000

a. Abhängige Variable: Einkommen in EUR

BLUE-Annahmen: Linearität

- Wenn zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen ein **nichtlinearer Zusammenhang** besteht, ist das lineare Regressionsmodell durch eine Transformation der unabhängigen Variablen anzupassen
- Es gibt verschiedene Formen nichtlinearer Zusammenhänge (z.B. u-förmig, glockenförmig, exponentiell, Sprungstelle), die jedoch theoretisch begründet werden sollten
- Im Beispiel vermuten wird, dass der Zusammenhang zwischen Berufserfahrung und Einkommen nicht linear, sondern glockenförmig ist

BLUE-Annahmen: Linearität

- Um diese Hypothese zunächst grafisch zu testen, betrachten wir ein Streudiagramm (y-Achse: Einkommen, x-Achse: Berufserfahrung) und lassen hier eine nichtparametrische Regressionslinie (Loess, Kernel-Regression) einzeichnen
- Tipp: Falls die abhängige und/oder unabhängige Variable relativ wenige Ausprägungen aufweist, empfiehlt es sich aus optischen Gründen, der entsprechenden Variable für das Streudiagramm einen Zufallsfehler (Jitter) zuzuspielen, hier z.B. für die Variable Berufserfahrung:

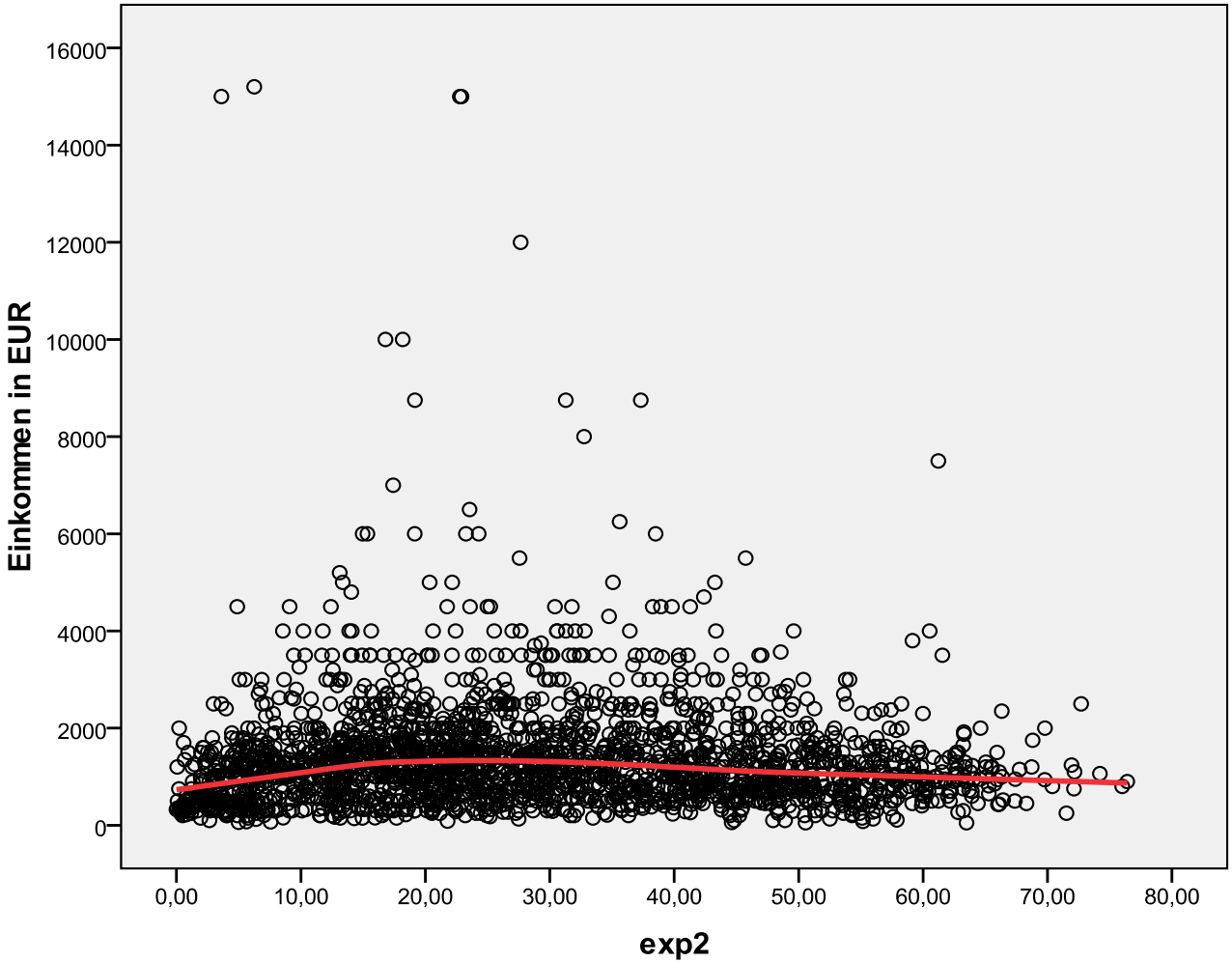
```
COMPUTE exp2 = exp + NORMAL(.5).
EXECUTE.
```

Zuspielen eines Zufallsfehlers

```
IF (exp2 < 0) exp2 = ABS(exp2).
EXECUTE.
```

Beibehalten der unteren Grenze 0

BLUE-Annahmen: Linearität



BLUE-Annahmen: Linearität

- Der Zusammenhang zwischen Einkommen und Berufserfahrung scheint erwartungsgemäß u-förmig zu sein
- Zum statistischen Test dieser Hypothese und zur Modifikation des Regressionsmodells gibt es mehrere Möglichkeiten:
  - (1.) Einteilen der Berufserfahrung in Abschnitte und Aufnahme entsprechender Dummy-Variablen in das Modell
  - (2.) Aufnahme eines quadrierten Terms für die Berufserfahrung (zusätzlich zum linearen Term) in das Modell
  - Variante 2 ist meist sparsamer und eleganter

BLUE-Annahmen: Linearität

- Vorgehensweise für Variante 2:
  - Ermittlung des arithmetischen Mittelwertes der Berufserfahrung (28,5 Jahre)
  - Zentrierung der Berufserfahrung (zur Vermeidung von Kollinearität zwischen dem linearen und quadrierten Term)
  - Quadrierung der zentrierten Berufserfahrung:

```
DESCRIPTIVES VARIABLES = exp.

COMPUTE exp_c = exp - 28.5.
EXECUTE.
COMPUTE exp_q = exp_c*exp_c.
EXECUTE.
```

BLUE-Annahmen: Linearität

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	-124,347	110,514		-1,125	,261
	Geschlecht: Frau	-542,324	43,528	-,229	-12,459	,000
	Wohnort: Ostdeutschland	-380,693	45,545	-,152	-8,359	,000
	Bildungsjahre	161,612	7,867	,396	20,544	,000
	Berufserfahrung zentriert	10,311	1,354	,151	7,617	,000
	Berufserfahrung quadriert	-,473	,070	-,130	-6,728	,000

a. Abhängige Variable: Einkommen in EUR

BLUE-Annahmen: Linearität

- Interpretationsrichtlinie für quadrierte Terme:
  - Ist der Effekt des quadrierten Terms negativ und signifikant (wie im Beispiel), handelt es sich um einen glockenförmigen Zusammenhang
  - Ist der Effekt des quadrierten Terms positiv und signifikant, handelt es sich um einen u-förmigen Zusammenhang

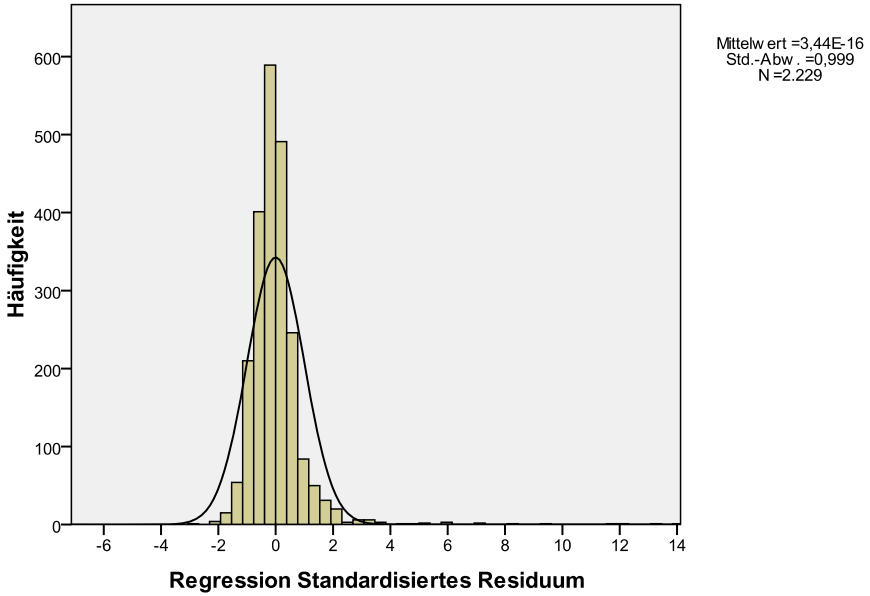
BLUE-Annahmen: Residuendiagnostik

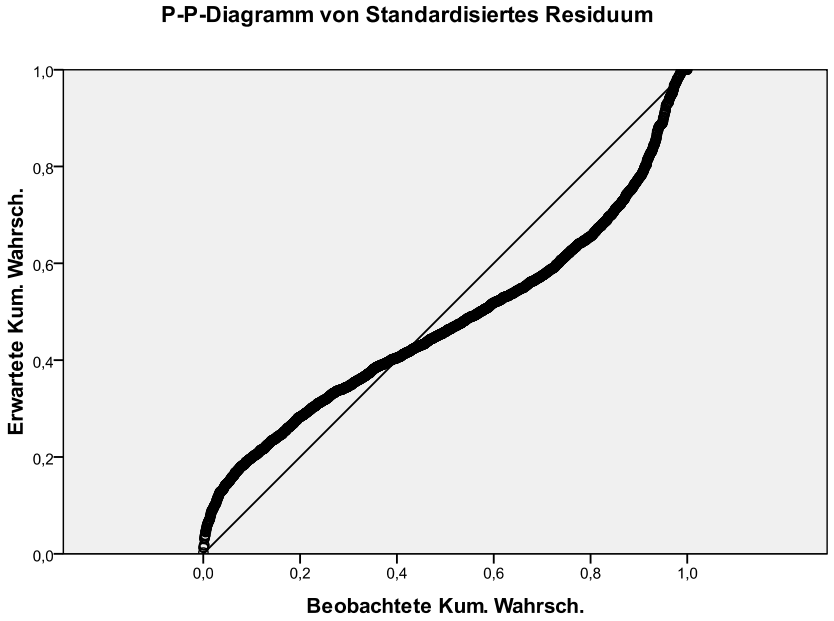
- Die **Residuen**, also die Abweichungen zwischen Beobachtungs- und Vorhersagewerten, sollten zufällig auftreten und keinem systematischen Muster folgen
- Andernfalls sind die Signifikanztests (F-Test, t-Tests) verzerrt
- Mögliche Ursachen für nicht-zufällige Residuen:
  - Wichtige Erklärungsgrößen fehlen im Modell
  - Es gibt Abhängigkeiten in den Daten (z.B. Klumpeneffekte)
  - Nichtlineare Zusammenhänge wurden nicht erkannt und modelliert
  - Die abhängige Variable ist schief verteilt

BLUE-Annahmen: Residuendiagnostik

- Zunächst prüfen wir, ob die Residuen symmetrisch verteilt sind
- Dazu wählen wir im Regressionsmenü unter „Diagramme“ das Histogramm und Normalverteilungsdiagramm (P-P-Diagramm) aus
- Wie im Histogramm ersichtlich ist, sind die Residuen tendenziell linkssteil verteilt
- Im P-P-Diagramm sind die Residuen dann normalverteilt, wenn die dicke Linie auf der dünnen Referenzlinie liegt
- Auch hier gibt es deutliche Abweichungen der Verteilung von einer Normalverteilung

BLUE-Annahmen: Residuendiagnostik

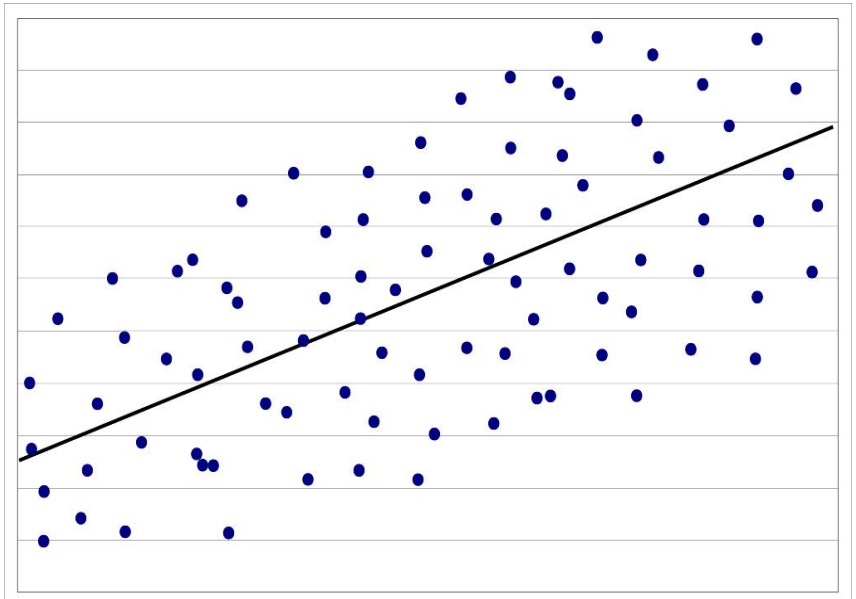




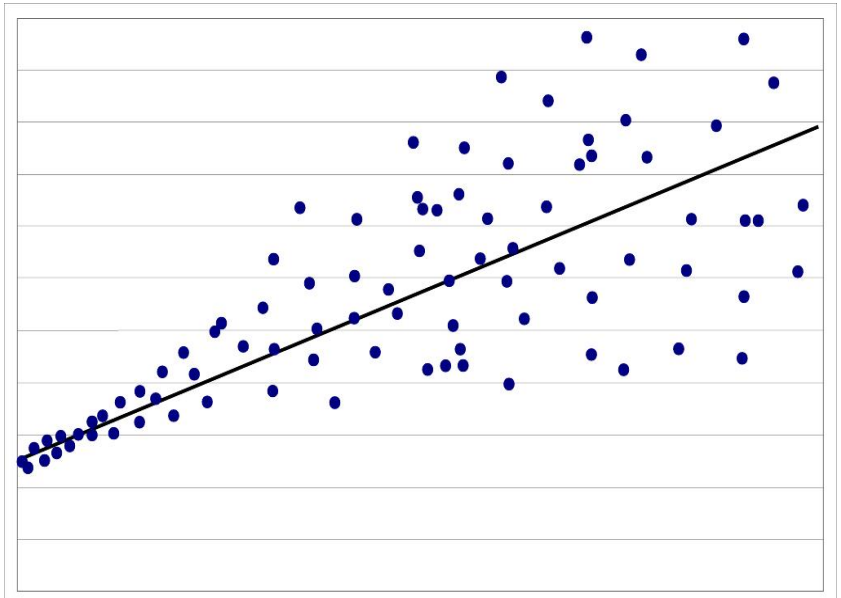
BLUE-Annahmen: Residuendiagnostik

- Weiterhin wird gefordert, dass eine Varianzengleichheit (Homoskedastizität) der Residuen gegeben sein sollte (Diagramm links nächste Folie)
- Unterscheiden sich die Residualvarianzen bei unterschiedlichen Ausprägungen der Variablen x, liegt Heteroskedastizität der Residuen vor (rechtes Diagramm)
- Bei ungleichen Residualvarianzen führt die OLS-Methode nicht zu effizienten Schätzwerten für die Regressionskoeffizienten
- D.h., dass diese Schätzwerte nicht die kleinst mögliche Varianz aufweisen; auch die t-Werte sind keine zuverlässigen Schätzer mehr

BLUE-Annahmen: Residuendiagnostik



Homoskedastizität



Heteroskedastizität

BLUE-Annahmen: Residuendiagnostik

- Typisches Beispiel für das Auftreten von Heteroskedastizität: bei einer Zeitreihe steigen die Abweichungen von der Trendgeraden mit Fortlauf der Zeit (z.B. für die Treffgenauigkeit bei der Wettervorhersage: je weiter in der Zukunft, desto unwahrscheinlicher ist eine genaue Prognose)
- Ob Varianzhomogenität vorliegt, kann durch einen Plot der Vorhersagefehler bzw. Residuen (y-Achse) gegen die Vorgersagewerte (x-Achse) beurteilt werden
- Dieser Plot ist jedoch häufig wenig aufschlussreich, weshalb hier eine andere Vorgehensweise empfohlen wird:

BLUE-Annahmen: Residuendiagnostik

- Test auf Homoskedastizität mithilfe von Box-Plots:
  - Speichern der standardisierten Residuen und der Vorhersagewerte als neue Variablen im Datensatz
  - Einteilung der Vorhersagewerte in Quartile
  - Box-Plot der standardisierten Residuen für die Quartile

BLUE-Annahmen: Residuendiagnostik

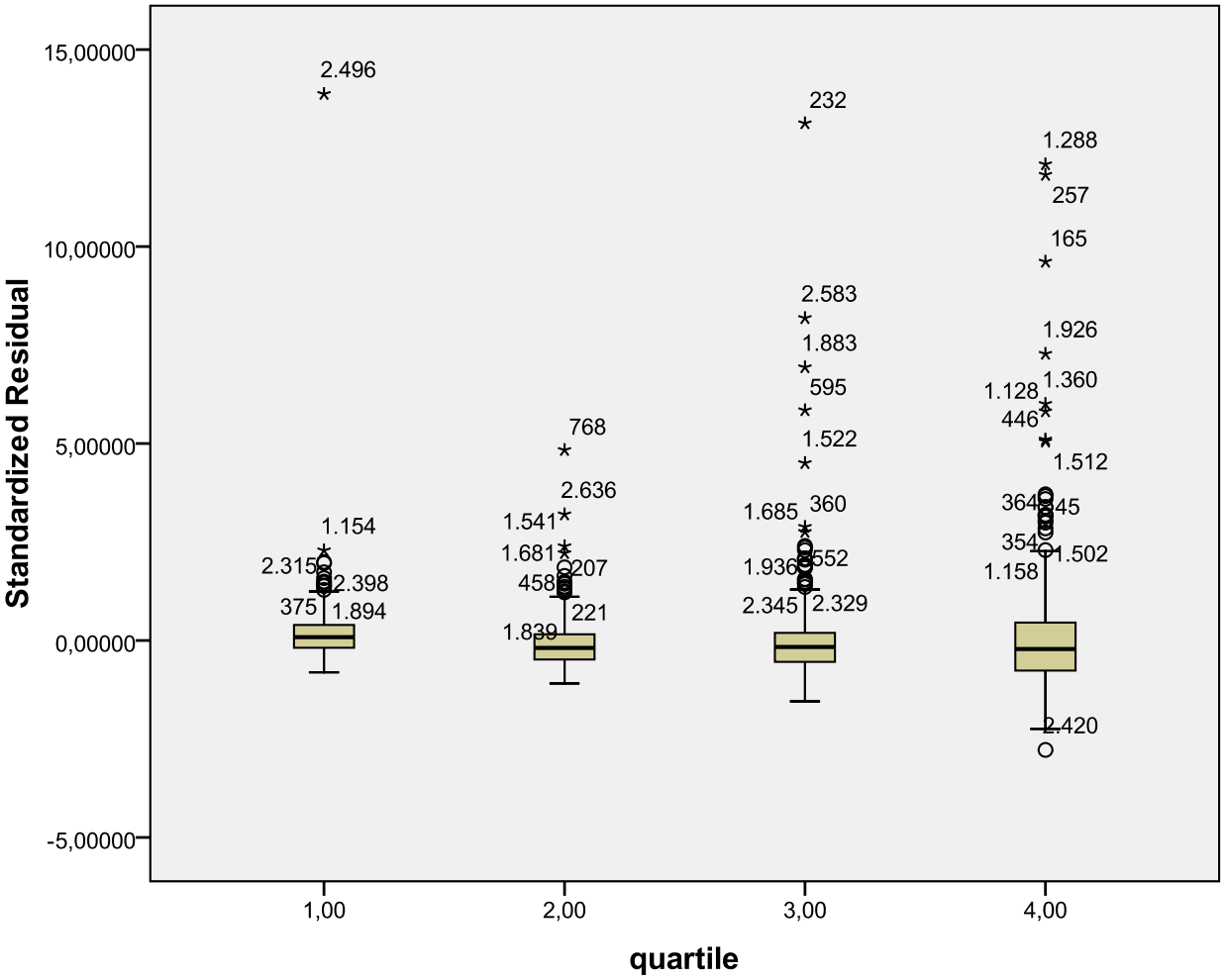
```
REGRESSION
  /DEPENDENT eink
  /METHOD=ENTER frau ost bild exp_c exp_q
  /SAVE PRED ZRESID.

FREQUENCIES VARIABLES=PRE_1
  /FORMAT=NOTABLE
  /NTILES=4
  /ORDER=ANALYSIS.

RECODE PRE_1 (lo thru 930.7 = 1) (930.71 thru 1247.6 = 2) (1247.61 thru
1703 = 3) (1703.1 thru hi = 4) INTO quartile.
EXECUTE.

EXAMINE VARIABLES=ZRE_1 BY quartile
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```

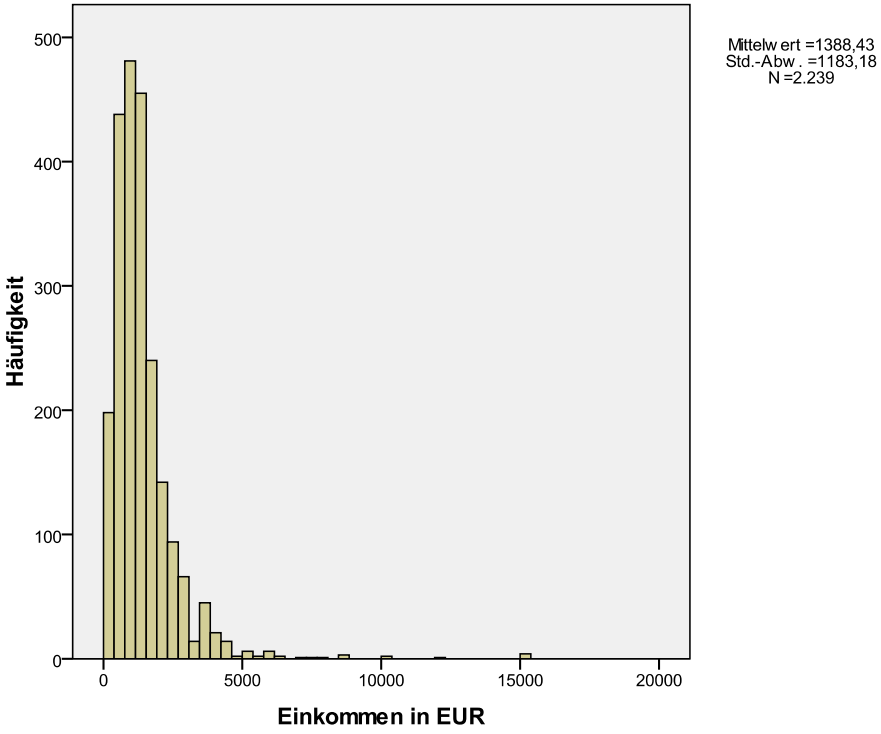
BLUE-Annahmen: Residuendiagnostik

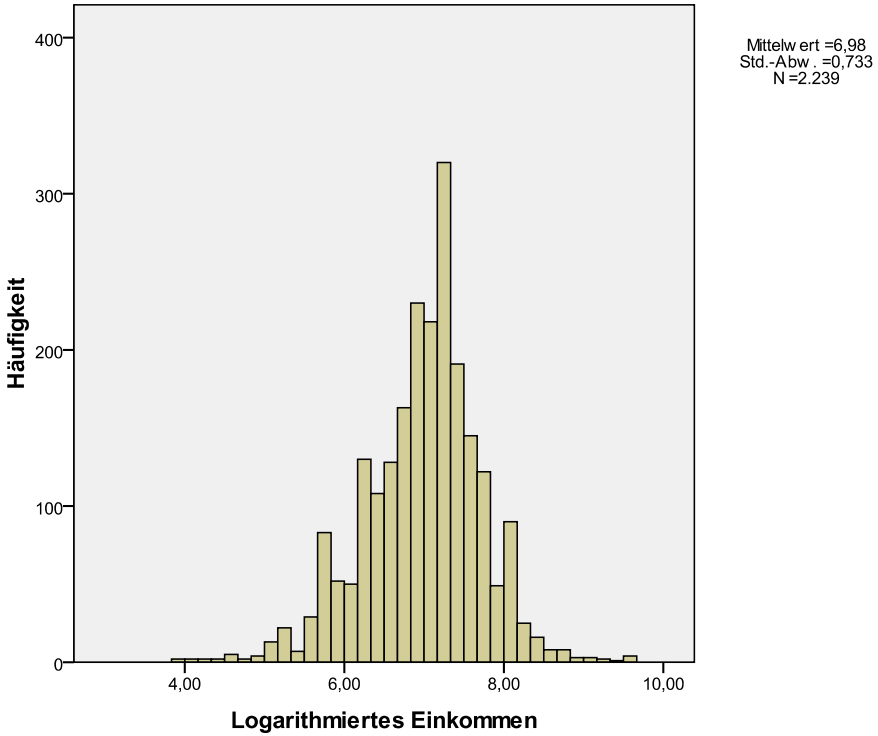


BLUE-Annahmen: Residuendiagnostik

- Der Box-Plot zeigt recht eindeutig, dass die Varianz der Residuen mit steigenden Vorhersagewerten (also im höheren Einkommensbereich) zunimmt, es liegt Heteroskedastizität vor
- Wie kann nun Abhilfe geschaffen werden, um die Probleme (schiefe Verteilung und Heteroskedastizität der Residuen) zu beheben?
- Wir vermuten, dass die Ursache der Probleme die typischerweise schiefe (linkssteile) Verteilung der abhängigen Variablen ist (nächste Folie, links)
- Wir nehmen daher eine Transformation der AV vor, indem wir das Einkommen logarithmieren, wodurch die Verteilung symmetrisch wird (rechts)

BLUE-Annahmen: Residuendiagnostik





BLUE-Annahmen: Residuendiagnostik

- Betrachten wir nun erneut ein Histogramm der Residuen, ein Normalverteilungsdiagramm der Residuen und den zuvor dargestellten Box-Plot (nächste Folien) zeigt sich, dass
  - Die Verteilung der Residuen nun annähernd symmetrisch ist
  - Sich auch im Normalverteilungsdiagramm kaum noch Abweichungen von der Referenzlinie zeigen
  - Die Varianz der Residuen über die Vorhersagewerte nun annähernd gleich ist (Homoskedastizität)

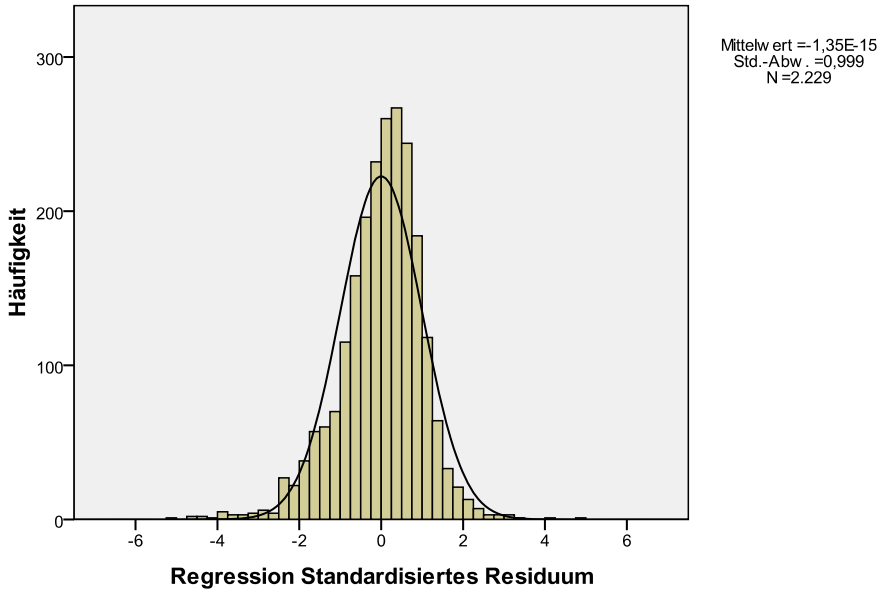
BLUE-Annahmen: Residuendiagnostik

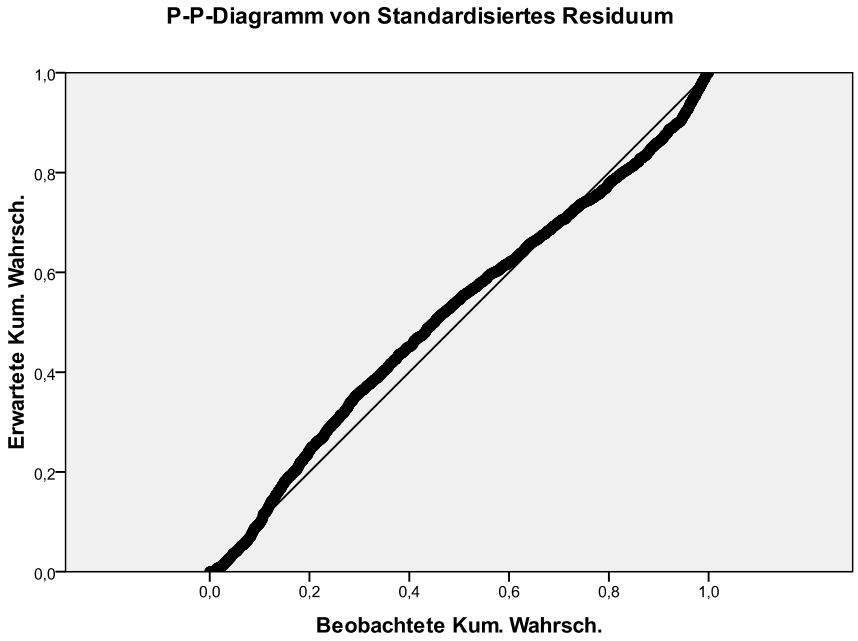
Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	6,102	,065		93,686	,000
	Geschlecht: Frau	-,430	,026	-,295	-16,759	,000
	Wohnort: Ostdeutschland	-,210	,027	-,137	-7,818	,000
	Bildungsjahre	,101	,005	,402	21,772	,000
	Berufserfahrung zentriert	,009	,001	,217	11,381	,000
	Berufserfahrung quadriert	,000	,000	-,177	-9,545	,000

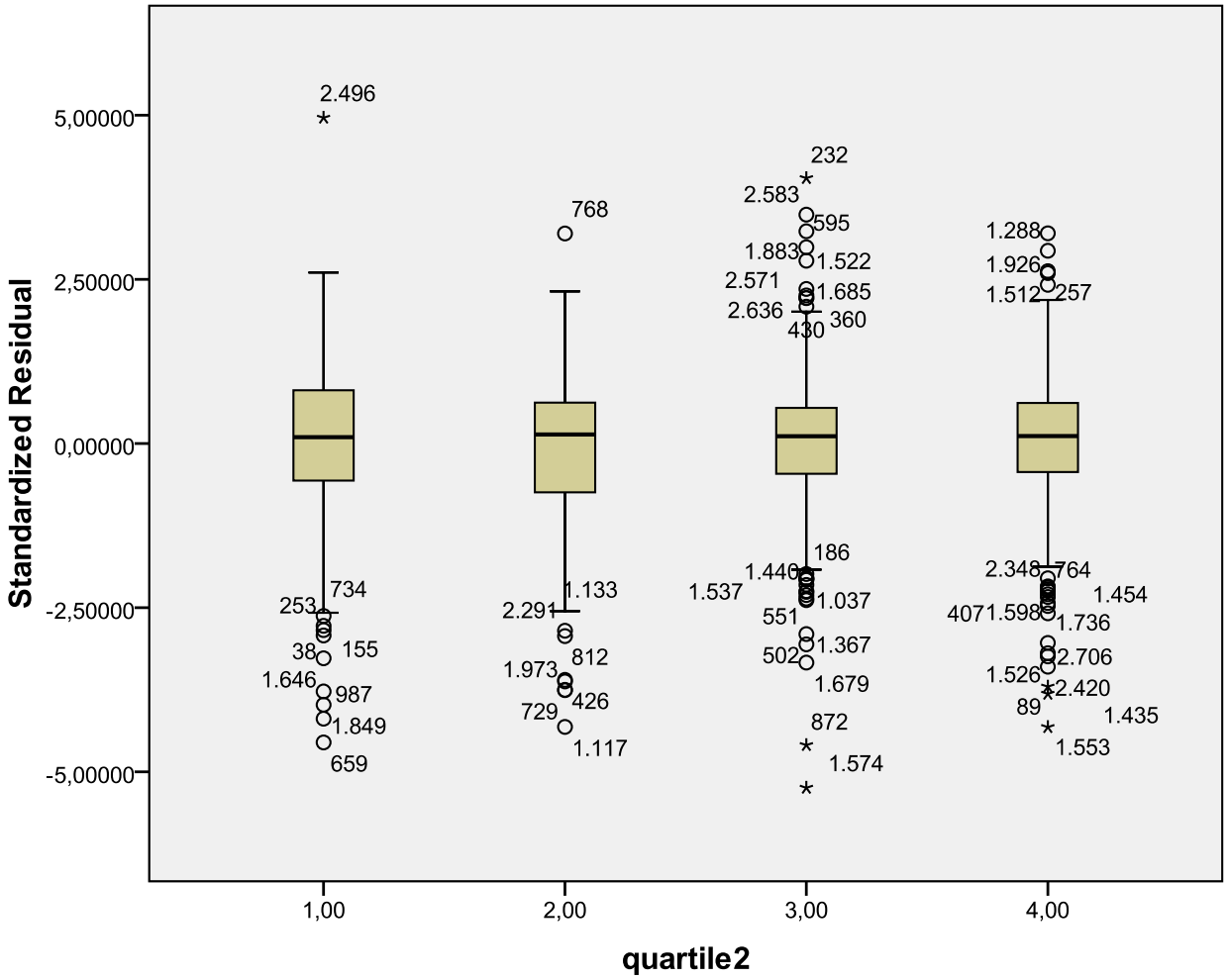
a. Abhängige Variable: Logarithmiertes Einkommen

BLUE-Annahmen: Residuendiagnostik





BLUE-Annahmen: Residuendiagnostik



BLUE-Annahmen: Kollinearität

- **Kollinearität (bzw. Multikollinearität)** liegt vor, wenn zwei oder mehrere unabhängige Variable sehr hoch miteinander korrelieren
- Bei perfekter Kollinearität ließe sich eine erklärende Variable über eine lineare Gleichung aus einer oder mehreren anderen erklärenden Variablen exakt berechnen
- Beispiel: In ein Regressionsmodell fließen die drei Variablen Partnerschaftsdauer zum Befragungsjahr, Jahr des Beginns der Partnerschaft und Befragungsjahr ein
- Die Partnerschaftsdauer ist nun nichts anderes als Befragungsjahr minus Jahr des Beginns der Partnerschaft und damit redundant

BLUE-Annahmen: Kollinearität

- Wenn zwar keine perfekte, aber eine hohe Kollinearität zwischen zwei Variablen besteht, können folgende Probleme auftreten:
  - Das
  - Das

BLUE-Annahmen: Kollinearität

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	4,684	,104		45,014	,000
	Berufserfahrung	-,107	,005	-2,558	-22,157	,000
	Alter in Jahren	,114	,005	2,615	22,652	,000

a. Abhängige Variable: Logarithmiertes Einkommen

Koeffizienten[a]

Modell		Kollinearitätsstatistik	
		Toleranz	VIF
1	(Konstante)		
	Berufserfahrung	,027	36,521
	Alter in Jahren	,027	36,521

a. Abhängige Variable: Logarithmiertes Einkommen

BLUE-Annahmen: Kollinearität

- Alter und Berufserfahrung korrelieren mit r = 0,987, werden aber trotzdem gemeinsam in ein lineares Regressionsmodell aufgenommen
- Dass dieses Modell Kollinearitätsprobleme hat, kann durch Toleranz und Varianzinflationsfaktor beurteilt werden
- Der Toleranzwert einer unabhängigen Variablen j ist definiert als:

$$\text{Toleranz}_j = 1 - R^2_j$$

- Dabei bezeichnet R^2_j die multiple quadrierte Korrelation der unabhängigen Variablen j mit den anderen unabhängigen Variablen des Modells
- Berechnung im Beispiel: 1-(0,987*0,987) = 0,027

BLUE-Annahmen: Kollinearität

- Der zusätzlich ausgegebene Varianzinflationsfaktor (VIF) ist nichts anderes als der Kehrwert der Toleranz (hier mit Rundung):

$$VIF_j = \frac{1}{Tol_j} = \frac{1}{1 - R^2_j} = \frac{1}{0{,}027} = 37{,}0$$

- Faustregel für die Interpretation: Toleranzwerte unter 0,1 oder VIF-Werte über 10 wecken den Verdacht auf Kollinearität; Toleranzwerte unter 0,01 lassen sicher auf das Vorliegen von Kollinearität schließen
- Abhilfe: Prädiktoren aus der Regression entfernen oder kollineare Prädiktoren durch Mittelwertbildung zu Skala zusammenfassen

Ausblick

- Zur Logik multivariater Regressionen mit mehr als einer unabhängigen Variablen siehe (→ „Forschungsdesigns und Drittvariablenkontrolle“)
- Das Verständnis der linearen Regression ist essentiell für die Einarbeitung in Erweiterungen einfacher linearer Regressionsmodelle wie die logistische Regressionen (verallgemeinertes lineares Modell) oder Mehrebenenmodelle (hierarchisches lineares Modell)
- Lineare Regression und Varianzanalyse basieren beide auf dem Allgemeinen Linearen Modell

Ausgewählte Literatur

- Allison, P. D. (1999): Multiple Regression. A Primer. Thousand Oaks: Pine Forge Press.

- Backhaus et al. (2011): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. X. Auflage. Berlin: Springer (Kapitel 1).

- Kopp, J. & Lois, D. (2014): Sozialwissenschaftliche Datenanalyse. Eine Einführung. 2. Auflage. Wiesbaden: Springer VS (Kapitel 5).

- Urban, D. & Mayerl, J. (2008): Regressionsanalyse: Theorie, Technik und Anwendung. X. Auflage Wiesbaden: VS.