17 Abh□ngigkeiten zwischen zwei Variablen: Einfache lineare Regression

Was Sie in diesem Kapitel lernen

- Kann man aus dem Vorbereitungsaufwand, den Studierende für eine Klausur erbringen, die Klausurnote vorhersagen?
- Wie kann man die Genauigkeit einer solchen Vorhersage exakt beziffern?
- Kann man auch die Vorhersagefehler genau beziffern und, wenn ja, wie?
- Wie hängt die Vorhersagegleichung von der Maßeinheit ab, in der die beteiligten Variablen vorliegen?
- Welchen Einfluss auf die Vorhersagegleichung hat die Standardisierung der beteiligten Variablen?
- Wie hängt der Korrelationskoeffizient mit dem Regressionsgewicht zusammen?

Im letzten Kapitel haben wir anhand einiger Beispiele erläutert, dass man von einem Zusammenhang zwischen zwei Variablen spricht, wenn die eine Variable aus der anderen vorhergesagt werden kann. Wir haben festgestellt, dass die Vorhersage umso genauer gelingt, je höher zwei Variablen miteinander korrelieren. Anschließend haben wir für verschiedene Kombinationen von Skalen Korrelationskoeffizienten eingeführt, die den Zusammenhang zwischen zwei Variablen mathematisch beschreiben.

Nun wollen wir uns etwas genauer damit befassen, wie die Messwerte einer Variablen aus denen der anderen Variablen vorhergesagt werden können. Wir beschränken uns dabei auf den Fall zweier metrischer Variablen. Betrachten wir zum Einstieg in die Aufgabenstellung noch einmal die Darstellung von Messwertpaaren in einem Punktediagramm. Wenn der Punkteschwarm, so wie in Abbildung 16.2 b (in Abschn. 16.2), eine Linie darstellt, können die Messwerte einer Variablen Y fehlerfrei aus den Messwerten der anderen Variablen X vorhergesagt werden. In diesem hypothetischen Fall beträgt die Produkt-Moment-Korrelation $r_{XY} = 1$. Dieser Fall ist z. B. gegeben, wenn man eine Variable mit sich selbst korreliert, die Messwerte der Merkmalsträger auf einer Variablen also einfach kopiert und die Korrelation der Variablen mit ihrer Kopie berechnet. Da die Messwerte jedes Messwertepaares identisch sind, kann fehlerfrei vom ersten Messwert eines Merkmalsträgers auf den zweiten geschlossen werden.

Mit Ausnahme der Korrelation einer Variablen mit sich selbst kommen perfekte Korrelationen in der Psychologie praktisch nicht vor. Bivariate Punkteschwärme entsprechen also praktisch nie Linien, sondern ähneln den Punkteschwärmen in Abbildung 16.2 a oder 16.2 c. Wie kann nun in diesen Fällen einer unvollständigen Korrelation eine Variable Y aus einer Variablen X vorhergesagt werden? Wie würde man z. B. vorgehen, wenn man bei einer Gruppe von Psychologiestudierenden die Punktezahl in der Methodenklausur (Y) aus der Stundenanzahl für die Klausurvorbereitung (X) vorhersagen wollte? Da die Korrelation zwischen diesen beiden Variablen mit Sicherheit nicht perfekt ist, wird uns eine exakte Vorhersage nicht gelingen.

Wenn wir in einem solchen Fall eine möglichst genaue Vorhersage treffen wollen, benötigen wir offensichtlich eine Methode, die eine Minimierung des Vorhersagefehlers gewährleistet. Die Regressionsanalyse ist eine solche Methode. »Regression« (von lat. regredi = zurückgehen) bedeutet in diesem Zusammenhang, dass eine abhängige Variable (AV) auf eine unabhängige Variable (UV) zurückgeführt wird. Eine UV wird auch als Prädiktor und eine AV als Kriterium bezeichnet (s. Abschn. 4.1.3). Eine Prädiktorvariable wird üblicherweise mit X, eine Kriteriumsvariable mit Y symbolisiert. Einen Prädiktor nennt man auch *Regressor* (Variable, auf die zurückgeführt wird) und ein Kriterium *Regressand* (Variable, die zurückgeführt wird). Will man anhand der Unterschiede in der UV Unterschiede in der AV erklären, spricht man auch von erklärender Variable (UV) und zu erklärender Variable (AV).

Eigenschaften des Mittelwerts

Um zu verstehen, wie die Regressionsmethode funktioniert, wollen wir uns die Eigenschaften des Mittelwertes in Erinnerung rufen. Nehmen wir einmal an, wir

wollten die Punktzahl in der Methodenklausur einer Psychologiestudentin prognostizieren, deren Vorbereitungsaufwand wir nicht kennen. Wie wir im letzten Kapitel gezeigt und am Beispiel »Geschlecht und Motorleistung« (in Abschn. 16.1) erläutert haben, wäre es in dieser Situation vernünftig, die Durchschnittspunktzahl der Methodenklausur ($\bar{y}$) anzugeben, weil der Mittelwert, sofern die Variable glockenförmig verteilt ist, derjenige Wert ist, der eine Verteilung am besten repräsentiert.

In Kapitel 6 hatten wir noch zwei weitere Eigenschaften des Mittelwertes kennengelernt, die für unsere Problemstellung von Bedeutung sind: Erstens beträgt die Summe aller Differenzen zwischen den Messwerten und dem Mittelwert (also die Summe aller $(y_m - \bar{y})$) immer 0, und zweitens ist die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert (also die Summe aller $(y_m - \bar{y})^2$) immer kleiner als die Summe der quadrierten Abweichungen der Messwerte von irgendeinem anderen Wert. Daraus folgt, dass wir mit dem Mittelwert denjenigen Wert prognostizieren, der den tatsächlichen Werten insgesamt am nächsten kommt. Wir machen also insgesamt den geringsten Vorhersagefehler, wenn wir zur Vorhersage der Methodennote einer beliebigen Studentin die Durchschnittsnote in Methodenlehre angeben. Diesen Durchschnitt bezeichnet man auch als *unbedingten Mittelwert*. Er ist »unbedingt«, da es nur einen gibt und dieser nicht von X abhängt.

Bedingte Mittelwerte

Nun kennen wir in unserem Gedankenexperiment zwar den Vorbereitungsaufwand der Psychologiestudentin nicht, wir wissen aber, dass es sich um eine Frau handelt. Wenn es nun so wäre, dass Studentinnen in Methodenklausuren typischerweise besser oder schlechter abschneiden würden als Studenten, könnten wir mit dieser Information die Genauigkeit unserer Prognose steigern. Analog zu unserem Beispiel »Geschlecht und Motorleistung« würden wir zur Prognose der Punktzahl unserer Studentin nicht mehr die Durchschnittsnote aller Studierenden heranziehen, sondern die Durchschnittsnote von Studentinnen.

Dieses Vorgehen wäre gleichbedeutend mit der Prognose der Punktzahl aus dem Geschlecht. Die Prädiktorvariable »Geschlecht« (X) kann zwei Werte annehmen (z. B. weiblich: $x = 0$; männlich: $x = 1$) und für jeden dieser beiden Werte lässt sich der Mittelwert der Kriteriumsvariablen bestimmen. Diese Mittelwerte nennt man bedingte Mittelwerte ($\bar{y}\,|\,x$). Unter der Bedingung, dass wir die Punktzahl einer Student*in* vorhersagen wollen, nehmen wir einen anderen Mittelwert als unter der Bedingung, dass wir die Punktzahl eines Studen*ten* vorhersagen wollen. Mithilfe der Zusatzinformation, dass Studentinnen und Studenten sich hinsichtlich ihrer Punktzahl in der Methodenklausur unterscheiden, können wir also unseren Vorhersagefehler verringern.

Die Genauigkeitssteigerung rührt daher, dass die Punktzahl innerhalb der beiden Geschlechtsgruppen weniger variiert als in der Gesamtgruppe von Männern und Frauen. Deshalb ist die Summe der quadrierten Abweichungswerte innerhalb der Geschlechtsgruppen kleiner als die Summe der quadrierten Abweichungswerte über alle Personen hinweg.

Lineare Regression

Nun übertragen wir dieses Prinzip auf eine metrische X-Variable und nehmen statt des Geschlechts die Anzahl der Vorbereitungsstunden als Prädiktor der Punktzahl in der Methodenklausur. Es ist anzunehmen, dass mit der Anzahl der Stunden, die jemand in die Vorbereitung auf die Klausur investiert, auch die Anzahl der in der Klausur erzielten Punkte steigt. Das würde bedeuten, dass X (Vorbereitungszeit) und Y (Punktzahl) positiv miteinander korreliert sind: $r_{XY} > 0$. Der beste Schätzwert für die erreichte Punktzahl einer Person, deren Vorbereitungszeit $x = 40$ Stunden betrug, wäre dann der Mittelwert aller Studierenden, deren Vorbereitungszeit ebenfalls $x = 40$ Stunden betrug. Statt des Geschlechts haben wir nun die Vorbereitungszeit als Bedingung, unter der wir den Mittelwert der erreichten Punktzahl bilden und als besten Schätzwert der individuellen Note angeben. Da die Vorbereitungszeit theoretisch in unendlich vielen Stufen variieren kann, haben wir es, anders als beim Geschlecht, nicht mehr nur mit zwei bedingten Mittelwerten zu tun, sondern theoretisch mit unendlich vielen. Praktisch lassen sich genau so viele bedingte Mittelwerte ermitteln, wie es Ausprägungen der unabhängigen Variablen gibt.

Analog zum Geschlechtsunterschied würde die Genauigkeit unserer Notenprognose dadurch zunehmen, dass innerhalb einer Gruppe von Personen mit gleicher Vorbereitungszeit die erzielten Klausurpunkte weniger stark streuen als in der Gesamtgruppe. Deshalb liegt der bedingte Mittelwert der Punktzahl (durchschnittliche

Punktzahl innerhalb einer Gruppe mit gleicher Vorbereitungszeit) näher an den tatsächlichen Punktzahlen der einzelnen Studierenden als der unbedingte Mittelwert (durchschnittliche Punktzahl aller Studierenden).

Ideal wäre es, wenn die bedingten Mittelwerte anhand einer einfachen Funktion der unabhängigen Variablen vorhergesagt werden könnten. Wenn alle bedingten Mittelwerte z. B. auf einer Geraden liegen würden, müsste man nur noch die Geradengleichung kennen, um anhand eines x-Wertes den y-Wert einer Person zu prognostizieren. Man müsste dann nicht mehr für jeden einzelnen x-Wert den bedingten Mittelwert tabellieren. Dies würde insbesondere bei vielen x-Werten die Prognose erleichtern. Im Rahmen der Regressionsanalyse versucht man, den Zusammenhang zwischen den bedingten Mittelwerten und der unabhängigen Variablen durch eine solche Funktion zu beschreiben.

Im Modell der einfachen linearen Regression wird angenommen, dass der Zusammenhang zwischen den bedingten Mittelwerten und den Werten der unabhängigen Variablen durch eine lineare Beziehung beschrieben wird. Im Idealfall liegen die bedingten Mittelwerte auf einer geraden Linie. Diese Gerade nennt man Regressionsgerade. Wenn der Zusammenhang zwischen den Variablen linear ist, lassen sich die y-Werte der Merkmalsträger anhand der Regressionsgeraden optimal aus ihren x-Werten vorhersagen.

Stichprobe und Population

In einer Stichprobe von Personen werden die bedingten Mittelwerte nicht perfekt auf einer Geraden liegen, auch wenn in der zugrunde liegenden Population das lineare Regressionsmodell gelten würde. Abweichungen der beobachteten von den geschätzten Werten werden schon allein durch die Stichprobenziehung und den damit verbundenen Stichprobenfehler zustande kommen (s. Abschn. 8.4). In der Population beschreibt die Regression die Abhängigkeit der bedingten Erwartung einer Variablen Y von der Variablen X (Steyer, 2003). Für jeden Wert x von X kann man den bedingten Erwartungswert von Y betrachten. Die bedingte Erwartung kann, muss aber nicht einer mathematisch einfach beschreibbaren Funktion folgen. In diesem Kapitel werden wir uns im ersten Teil auf die Deskriptivstatistik der einfachen linearen Regressionsanalyse beschränken, das theoretische Modell der bedingten Erwartung sowie inferenzstatistische Fragen der einfachen Regressionsanalyse werden wir in Abschnitt 17.9 behandeln.

Ziel des vorliegenden Kapitels ist es zu zeigen, wie eine angenommene lineare Beziehung zwischen zwei Variablen beschrieben werden kann. Wir gehen daher von konkreten x- und y-Werten aus und zeigen, wie man in einen vorhandenen Punkteschwarm eine Gerade optimal einpassen kann. Die Grundidee der Regressionsanalyse werden wir weiterhin am Beispiel der Prädiktion erläutern.

Wahrer und approximativer linearer Zusammenhang

Die Annahme, dass der Zusammenhang zwischen der abhängigen und der unabhängigen Variablen linearer Natur ist, kann falsch sein. Wenn man beispielsweise die Körpergröße aus dem Alter vorhersagen wollte, wäre es sicher falsch, davon auszugehen, dass der Zusammenhang linear ist. Denn die Körpergröße nimmt zwar bis zu einem bestimmten Alter stetig zu, sie bleibt aber, nachdem Menschen ausgewachsen sind, über längere Zeit konstant und nimmt im Alter allmählich wieder ab. Die altersbedingten Größenmittelwerte liegen folglich nicht auf einer Geraden, sondern auf einer gekrümmten Linie. Man spricht deshalb auch von einem kurvilinearen Zusammenhang (vgl. auch Abb. 16.2 d in Abschn. 16.2).

Obwohl die Annahme der Linearität nicht immer stimmt, wird das Modell der linearen Regression häufig zur Prognose herangezogen. Dies hat zwei Gründe: Erstens stehen psychologische Variablen sehr oft in einem annähernd linearen Zusammenhang. Zweitens sind die Abweichungen von der Linearität häufig so gering, dass sie als unsystematische Schwankungen interpretiert werden können. Sofern die Abweichungen von der Linearität unbedeutend sind, nimmt man einen geringfügig größeren Vorhersagefehler in Kauf, um sich die attraktiven Eigenschaften des linearen Regressionsmodells erhalten zu können, die in seiner Anschaulichkeit und in seiner einfachen mathematischen Formulierung bestehen.

17.1 Kleinste-Quadrate-Kriterium

Unabhängig davon, ob der Zusammenhang zwischen zwei Variablen linear ist oder nicht, stellt die Summe der quadrierten Differenzen zwischen den anhand der unabhängigen Variable X vorhergesagten Werten und den beobachteten y-Werten ein Kriterium für die Optimierung der Vorhersage mittels der Regressions-

methode bereit. Den anhand eines x_m-Wertes vorhergesagten y_m-Wert bezeichnen wir mit $\hat{y}_m$. Man nennt dieses Kriterium das Kleinste-Quadrate-Kriterium. Im Falle der linearen Regression schreibt es vor, die Regressionsgerade so in den Punkteschwarm zu legen, dass die Summe der quadrierten Abstände der beobachteten Kriteriumswerte von der Regressionsgeraden ein Minimum ergibt. Die Regressionsmethode minimiert also die Summe der Abweichungsquadrate (*SAQ*):

$$SAQ = \sum_{m=1}^{n} (y_m - \hat{y}_m)^2 \rightarrow \text{min!} \qquad \text{(F 17.1)}$$

Die Bedeutung der *SAQ* hatten wir bei der Definition der Varianz erstmals kennengelernt (vgl. Formel F 6.26). Der einzige Unterschied in der Bedeutung des Begriffs hier und dort besteht darin, dass wir es hier mit Abweichungen von geschätzten Werten ($\hat{y}_m$) zu tun haben, die man anhand eines Regressionsmodells erhält. Die Optimierung der Vorhersage anhand des Kleinste-Quadrate-Kriteriums ist also gleichbedeutend mit der Minimierung der quadrierten Abweichungen.

Wir wollen uns die lineare Regression anhand eines fiktiven Zahlenbeispiels verdeutlichen (s. folgenden Kasten). Betrachten wir zunächst den Punkteschwarm, der sich aus den Messwertpaaren der 25 Merkmalsträger ergibt. Er ist in Abbildung 17.1 dargestellt. Man sieht sofort, dass die Punkte nicht auf einer Linie liegen. Der Zusammenhang zwischen den beiden Variablen ist also nicht perfekt. Allerdings hat der Punkteschwarm eine gerade, stark gestreckte Form. Daraus können wir schließen, dass die beiden Variablen miteinander hoch korreliert sind. Außerdem sehen wir, dass der Punkteschwarm von links nach rechts ansteigt. Da die Abszisse von links nach rechts und die Ordinate von unten nach oben zunehmende Werte abbilden, können wir aus der Lage des Punkteschwarms eine positive Korrelation erschließen. Eine Berechnung der Korrelation bestätigt unseren Eindruck: Die Produkt-Moment-Korrelation beträgt $r_{XY} = 0{,}88$. Man kann also in diesem fiktiven Beispiel die in der Methodenklausur erzielten Punkte gut aus der Anzahl der für die Vorbereitung aufgewendeten Stunden vorhersagen.

Beispiel

Lernaufwand und erreichte Punktzahl in einer Klausur

In Tabelle 17.1 stehen die Messwerte beider Variablen von 25 namentlich bezeichneten Psychologiestudierenden. Die erste Wertespalte enthält die Messwerte der Studierenden auf der Prädiktorvariablen (Anzahl von Stunden x_m der Vorbereitung auf die Methodenklausur). In der nächsten Wertespalte sind die Punkte y_m der jeweiligen Person in der Methodenklausur aufgeführt. Unter den Messwertzeilen dieser beiden Spalten sind die Messwertsummen, die Mittelwerte der Variablen, die empirischen Standardabweichungen und die empirischen Varianzen notiert.

Tabelle 17.1 Datenbeispiel zur linearen Regression (*X*: Klausurvorbereitung in Stunden; *Y*: Klausurpunkte)

	m	x_m	y_m	$\hat{y}_m$	$e_m = y_m - \hat{y}_m$	$e_m^2 = (y_m - \hat{y}_m)^2$
Bauer	1	18	21	19	2	4
Bergmann	2	26	22	23	−1	1
Diener	3	46	37	33	4	16
Fischer	4	42	30	31	−1	1
Förster	5	20	19	20	−1	1
Fuhrmann	6	26	25	23	2	4
Gärtner	7	38	32	29	3	9
Schreiber	8	34	32	27	5	25
Köhler	9	40	30	30	0	0

Tabelle 17.1 (Fortsetzung)

	m	x_m	y_m	$\hat{y}_m$	$e_m = y_m - \hat{y}_m$	$e_m^2 = (y_m - \hat{y}_m)^2$
Küfer	10	30	22	25	−3	9
Maler	11	24	26	22	4	16
Müller	12	14	19	17	2	4
Richter	13	44	29	32	−3	9
Schäfer	14	10	13	15	−2	4
Schmied	15	28	27	24	3	9
Schneider	16	28	21	24	−3	9
Gerber	17	36	25	28	−3	9
Schuster	18	16	16	18	−2	4
Steiger	19	50	33	35	−2	4
Steinmetz	20	24	17	22	−5	25
Töpfer	21	36	28	28	0	0
Wagner	22	32	23	26	−3	9
Weber	23	34	26	27	−1	1
Weidner	24	22	23	21	2	4
Zöllner	25	32	29	26	3	9
Summe		750,00	625,00	625,00	0,00	186,00
Mittelwert		30,00	25,00	25,00	0,00	7,44
Standardabweichung		10,09	5,73	5,04	2,73	
Varianz		101,76	32,88	25,44	7,44	

Welchen $\hat{y}$-Wert werden wir beim Vorliegen eines bestimmten x-Wertes prognostizieren? Denjenigen Wert, den die Regressionsgerade dem jeweiligen x-Wert zuordnet: Einem x_m-Wert von 50 ordnet die Regressionsgerade den $\hat{y}_m$-Wert von 35 zu; einem x_m-Wert von 24 den $\hat{y}_m$-Wert 22 usw. Für eine Person, die sich $x_m = 50$ Stunden auf die Klausur vorbereitet hat, erwarten wir also $\hat{y}_m = 35$ Klausurpunkte; für eine Person, die sich $x_m = 24$ Stunden vorbereitet hat, erwarten wir $\hat{y}_m = 22$ Klausurpunkte. Die dritte Wertespalte von Tabelle 17.1 enthält die prognostizierten $\hat{y}_m$-Werte aller Merkmalsträger.

Abbildung 17.1 zeigt, dass diese Erwartungen falsch sein können und es überwiegend auch sind. Nur in zwei Fällen liegen die Messwerte exakt auf der Regressionsgeraden. Nur in diesen beiden Fällen stimmen also die vorhergesagten Werte mit den beobachteten überein. In allen anderen Fällen liegen die beobachteten Werte entweder über oder unter der Geraden, streuen also um den vorhergesagten Wert. Betrachten wir die Situation bei Psychologiestudentin Steinmetz etwas genauer. Sie hat sich $x_{20} = 24$ Stunden auf die Klausur vorbereitet. Die Regression lässt $\hat{y}_{20} = 22$ Klausurpunkte für sie erwarten. Tatsächlich erreichte Frau Steinmetz aber nur $y_{20} = 17$ Punkte. Ihre Leistung wurde also um $\hat{y}_{20} - y_{20} = 5$ Punkte überschätzt.

Abbildung 17.1 zeigt weiterhin, dass auch die bedingten y-Mittelwerte (also die Mittelwerte aller beobachteten y-Werte mit identischem x-Wert) meistens nicht genau auf der Regressionsgeraden liegen. Viele

x-Werte kommen nur einmal vor, charakterisieren also nur einen der 25 Studierenden. In diesen Fällen ist die durchschnittliche Punktzahl in der Klausur identisch mit der Punktzahl dieses einen Merkmalsträgers. Wenn sein y-Wert nicht auf der Regressionsgeraden liegt, liegt auch der bedingte y-Mittelwert nicht auf der Geraden. Aber auch bei mehreren Merkmalsträgern mit identischem x-Wert liegt der (bedingte) Mittelwert der y-Werte häufig nicht auf der Regressionsgeraden, so z. B. bei den beiden Personen Steinmetz und Maler. Beide haben $x_{11} = x_{20} = 24$ Stunden gelernt. Steinmetz hat $y_{20} = 17$ Punkte in der Klausur erreicht, Maler hingegen $y_{11} = 26$. Der Mittelwert aus beiden y-Werten beträgt $(\bar{y} \mid x = 24) = 21{,}5$. Der vorhergesagte Wert beträgt jedoch $(\hat{y} \mid x = 24) = 22$. Wir können daraus schließen, dass der Zusammenhang zwischen den beiden Variablen nicht-linear ist. Allerdings sind die Abweichungen von der Linearität sowohl optisch als auch numerisch relativ gering. Deshalb ist es vernünftig, mit dem Modell der linearen Regression zu operieren. Wenn man das tut, gibt es keine andere Gerade, die das Kleinste-Quadrate-Kriterium besser erfüllt als die in Abbildung 17.1 eingezeichnete Regressionsgerade.

Wie gut die Vorhersage der Punktzahl durch die Vorbereitungszeit ist, können wir der vierten und fünften Wertespalte von Tabelle 17.1 entnehmen. Dort sind die Differenzen zwischen den beobachteten und den vorhergesagten Klausurpunkten $y_m - \hat{y}_m$ sowie die Abweichungsquadrate $(y_m - \hat{y}_m)^2$ eingetragen. Die einfachen Differenzen werden mit e_m bezeichnet (der Buchstabe e rührt von engl. »error« = Fehler her, denn bei den Abweichungen handelt es sich um Prognose- oder Schätzfehler); die quadrierten Differenzen werden mit e_m^2 bezeichnet.

Beide Spalten verdienen Beachtung. Zunächst sieht man, dass die Summe der einfachen Differenzen e_m gleich 0 ist. Dieser Sachverhalt zeigt uns, dass die Regressionsgerade so gewählt wurde, dass es durchschnittlich (über alle Personen hinweg) zu keiner Verschätzung der Klausurpunkte kommt. Das sieht man auch daran, dass die Mittelwerte der y-Werte und der $\hat{y}$-Werte identisch sind: $\bar{y} = \bar{\hat{y}}$. Durchschnittlich werden $\bar{y} = 25$ Punkte erreicht, und diesen Mittelwert sagt auch die Regression vorher. Betrachten wir nun die Abweichungsquadrate e_m^2. Ihre Summe beträgt 186, und keine andere gerade Linie durch den Punkteschwarm würde zu einem kleineren Wert führen.

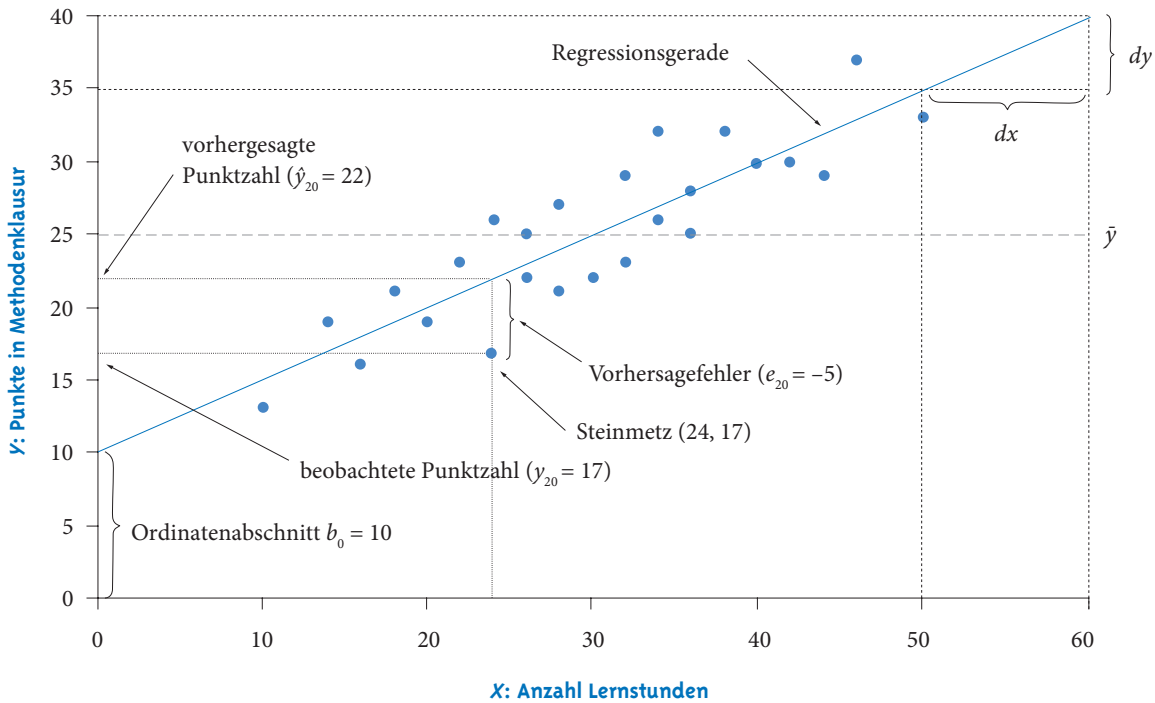

Abbildung 17.1 Einfache lineare Regression zum Zahlenbeispiel aus Tabelle 17.1

17.2 Regressionsgleichung

Mathematisch lässt sich die lineare Beziehung zwischen den vorhergesagten individuellen Werten $\hat{y}_m$ und den individuellen Werten x_m auf der unabhängigen Variablen wie folgt beschreiben:

$$\hat{y}_m = b_0 + b_1 \cdot x_m \qquad \text{(F 17.2)}$$

Insbesondere im nächsten und im übernächsten Kapitel werden wir auch auf folgende analoge Variablenschreibweise zurückgreifen:

$$\hat{Y} = b_0 + b_1 \cdot X \qquad \text{(F 17.3)}$$

Die Gleichung F 17.2 bezieht sich auf den Zusammenhang zwischen den Werten eines Merkmalsträgers m, Gleichung F 17.3 auf die Variablen, deren Werte die Merkmalsausprägungen sind.

Regressionsgewicht b_1. In diesen Gleichungen bezeichnet b_1 das Regressionsgewicht. Das Regressionsgewicht ist für alle Merkmalsträger gleich. Es wird über die erste Ableitung der Funktion bestimmt und beziffert die Steigung der Regressionsgeraden. Die Steigung der Geraden lässt erkennen, um wie viele Einheiten $\hat{Y}$ zunimmt, wenn X um eine Einheit zunimmt. Abbildung 17.1 veranschaulicht, was Steigung bedeutet und wie sie bestimmt werden kann. Man nimmt zwei hypothetische Messwertpaare, die auf der Regressionsgeraden liegen, und bildet für die beiden x-Werte ebenso wie für die beiden y-Werte die Differenz. Diese Differenzen sind, der Konvention entsprechend, in Abbildung 17.1 mit dx und dy bezeichnet. Der Quotient dy/dx gibt die Steigung der Regressionsgeraden an, also das Gewicht b_1, mit dem X zur optimalen Vorhersage von Y multipliziert wird. Der Koeffizient b_1 wird auch als Steigungskoeffizient oder in Anlehnung an den englischen Begriff für »Steigung« als *Slope* bezeichnet.

Achsenabschnitt b_0. Der Koeffizient b_0 wird auch als additive Konstante oder Achsen- oder Ordinatenabschnitt bezeichnet. Er ist der $\hat{y}$-Wert am Schnittpunkt der Regressionsgeraden mit der Ordinate und damit derjenige $\hat{y}$-Wert, den die Regressionsfunktion einem x-Wert von 0 zuordnet. Da die Regressionsgerade ein Stück der Ordinate abschneidet, wird b_0 in Anlehnung an den englischen Begriff für »Abschnitt« auch als *Intercept* bezeichnet.

Bestimmung der Regressionskoeffizienten

Wie gelangt man zu den beiden Regressionskoeffizienten b_0 und b_1? Es lässt sich zeigen, dass diejenige Gerade das Kleinste-Quadrate-Kriterium

$$SAQ = \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$
$$= \sum_{m=1}^{n} (y_m - (b_0 + b_1 \cdot x_m))^2 \rightarrow \min!$$

am besten erfüllt, deren Steigung b_1 in folgender Beziehung zum Produkt-Moment-Korrelationskoeffizienten steht:

$$b_1 = r_{XY} \cdot \frac{s_Y}{s_X} = \frac{s_{XY}}{s_X^2} \qquad \text{(F 17.4)}$$

Daraus folgt, dass bei z-standardisierten Variablen (s. Abschn. 6.5) das Regressionsgewicht mit dem Korrelationskoeffizienten identisch ist. Die Korrelation ist also nichts anderes als das Regressionsgewicht zweier z-standardisierter Variablen. Daraus können wir ersehen, dass die Produkt-Moment-Korrelation den linearen Zusammenhang zwischen zwei Variablen beschreibt. Die Korrelation zwischen X und Y beträgt in unserem Beispiel $r_{XY} = 0{,}88$. Wenn man diesen Wert mit dem Quotienten der Standardabweichungen von Y und X multipliziert (s. Tab. 17.1), ergibt sich ein Regressionsgewicht von $b_1 = 0{,}5$.

Dieses Regressionsgewicht bedeutet, dass $\hat{Y}$ um eine halbe Maßeinheit zunimmt, wenn X um eine ganze Maßeinheit zunimmt. Jede zusätzliche Stunde Klausurvorbereitung zahlt sich in unserem Beispiel also in einem erwarteten Gewinn von einem halben Klausurpunkt aus.

Kommen wir nun zum Ordinatenabschnitt b_0. Er wird wie folgt bestimmt:

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \qquad \text{(F 17.5)}$$

Bilden wir in unserem Beispiel nach Formel F 17.5 die Differenz zwischen dem Mittelwert von Y und dem mit b_1 gewichteten Mittelwert von X, ergibt sich: $b_0 = 25 - 0{,}5 \cdot 30 = 10$. Der Ordinatenabschnitt b_0 ist derjenige $\hat{y}$-Wert, den man erwartet, wenn $x = 0$ ist. Übertragen auf unser Beispiel bedeutet dies, dass Psychologiestudierende, die überhaupt nicht für die Klausur lernen (also einen Wert von $x = 0$ haben), mit $(\hat{y} \mid x = 0) = b_0 = 10$ Klausurpunkten rechnen können. Jede zusätzliche Stunde lässt einen halben Klausurpunkt mehr erwarten.

Während wir die vorhergesagten $\hat{y}$-Werte bisher anhand von Abbildung 17.1 geometrisch abgelesen hatten, können wir nun nach Kenntnis der Werte von b_0 und b_1 die unter x erwarteten $\hat{y}$-Werte auch algebraisch bestimmen. Hierzu greifen wir auf Gleichung F 17.2 zurück und setzen für b_0 den Wert 10 und für b_1 den Wert 0,5 sowie die interessierenden x_m-Werte ein und errechnen die zugehörigen $\hat{y}_m$-Werte. Dann ergibt sich die Gleichung

$$\hat{y}_m = 10 + 0{,}5 \cdot x_m.$$

17.3 Regressionsresiduum

Die Differenz $y_m - \hat{y}_m$ zwischen einem beobachteten und vorhergesagten y-Wert nennt man Regressionsresiduum (oder kurz: Residuum), Residualwert oder Fehlerwert e_m. Will man die y-Werte anhand der x-Werte vorhersagen, so zeigen die Residuen $e_m = y_m - \hat{y}_m$ den Vorhersagefehler an. Je größer die Fehlerwerte, umso größer ist die Abweichung eines beobachteten vom vorhergesagten Wert. Ist ein Fehlerwert gleich 0, liegt der beobachtete Wert auf der Regressionsgeraden. Sind alle Fehlerwerte gleich 0, liegen alle beobachteten Werte auf der Regressionsgeraden, die Vorhersage ist dann perfekt. Von Regressions*residuum* wird gesprochen, weil nach der bestmöglichen Vorhersage der y-Werte aus den x-Werten ein Rest der y-Werte zurückbleibt, der nicht vorhergesagt oder, wie man auch sagt, erklärt werden konnte.

Eine einfache Umformung zeigt, dass sich die y-Werte additiv aus den vorhergesagten $\hat{y}$-Werten und den Vorhersagefehlern zusammensetzen:

$$y_m = \hat{y}_m + e_m = b_0 + b_1 \cdot x_m + e_m \qquad \text{(F 17.6)}$$

In Variablenschreibweise:

$$Y = \hat{Y} + E = b_0 + b_1 \cdot X + E \qquad \text{(F 17.7)}$$

Vertiefung

Die **Residualwerte** weisen folgende Eigenschaften auf, die sich aufgrund der Kleinste-Quadrate-Schätzung ergeben:

(1) Die Summe aller Regressionsresiduen ist gleich 0:

$$\sum_{m=1}^{n} e_m = \sum_{m=1}^{n} (y_m - \hat{y}_m) = 0 \qquad \text{(F 17.8)}$$

(2) Die Summe aller quadrierten Regressionsresiduen ist minimal:

$$\sum_{m=1}^{n} e_m^2 = \sum_{m=1}^{n} (y_m - \hat{y}_m)^2 \rightarrow \text{min!} \qquad \text{(F 17.9)}$$

(3) Die Korrelation zwischen X und E ist gleich 0:

$$r_{XE} = 0 \qquad \text{(F 17.10)}$$

(4) Die Korrelation zwischen $\hat{Y}$ und E ist ebenfalls gleich 0:

$$r_{\hat{Y}E} = 0 \qquad \text{(F 17.11)}$$

Die erste Eigenschaft besagt, dass die Summe der Residualwerte 0 ist; d. h., über alle Werte hinweg betrachtet, mitteln sich die Fehler aus. Die zweite Eigenschaft ist das Kleinste-Quadrate-Kriterium, wodurch die Fehlerstreuung minimiert wird. Der dritten Eigenschaft zufolge sind die X-Variable und die Fehlervariable unkorreliert. Die Residualwerte repräsentieren somit den Teil des Merkmals Y, der nicht mit dem Merkmal X zusammenhängt. Da X und $\hat{Y}$ immer perfekt miteinander korreliert sind (denn schließlich wurden alle $\hat{y}$-Werte ja so vorhergesagt, dass sie genau auf einer Linie liegen), ist auch die Korrelation zwischen den erwarteten $\hat{y}$-Werten und den Regressionsresiduen gleich 0.

Residualvarianz und Standardschätzfehler

Alle Kennwerte der deskriptiven Statistik, die wir in Kapitel 6 kennengelernt haben, lassen sich auch für die Regressionsresiduen berechnen. Der Mittelwert der Regressionsresiduen beträgt 0. Die Varianz der Residualwerte gibt an, wie stark die beobachteten y-Werte um die Regressionsgerade und damit um die vorhergesagten $\hat{y}$-Werte streuen. Je größer bei einer gegebenen Maßeinheit von Y die Varianz der Residualwerte ausfällt, desto ungenauer war die Vorhersage, desto größer also der Vorhersagefehler. Deshalb bezeichnet man die Residualvarianz auch als Fehlervarianz. Ihre Bestimmungsgleichung lautet:

$$s_E^2 = \frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n} \qquad \text{(F 17.12)}$$

Wenn man aus der Fehlervarianz die Quadratwurzel zieht, erhält man die Standardabweichung des Regressionsresiduums, den sog. Standardschätzfehler:

$$s_E = \sqrt{\frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n}} \qquad \text{(F 17.13)}$$

Der Standardschätzfehler steht mit der Produkt-Moment-Korrelation in folgender Beziehung:

$$s_E = s_Y \cdot \sqrt{1 - r_{XY}^2} \qquad \text{(F 17.14)}$$

Je größer die Korrelation ist, umso geringer ist die Streuung der Residualwerte und umso geringer der Standardschätzfehler. Dies leuchtet unmittelbar ein, da wir in Kapitel 16 gesehen haben, dass die Datenpunkte umso enger um die Gerade liegen, je höher die Korrelation ist.

17.4 Quadratsummenzerlegung und Varianzzerlegung

Wie man sich leicht anhand von Abbildung 17.1 veranschaulichen kann, sind zwei Abweichungswerte besonders interessant. Der Abweichungswert $y_m - \bar{y}$ gibt an, wie stark ein beobachteter y-Wert vom (unbedingten) Mittelwert abweicht. Würde man die x-Werte nicht kennen, so wäre der Mittelwert $\bar{y}$ der beste vorhergesagte Wert für den y-Wert einer Person. Der Abweichungswert $y_m - \hat{y}_m$ zeigt die Abweichung des beobachteten Wertes von dem Wert an, den man aufgrund des x-Wertes prädizieren würde. In dem Maße, in dem sich der Absolutbetrag der Differenz $y_m - \hat{y}_m$ im Vergleich zum Absolutbetrag der Differenz $y_m - \bar{y}$ verringert, verbessert sich die Prognose durch die Hinzunahme der x-Werte. Beide Abweichungswerte lassen sich wie folgt in eine Beziehung bringen:

$$y_m - \bar{y} = (y_m - \hat{y}_m) + (\hat{y}_m - \bar{y}) \qquad \text{(F 17.15)}$$

Hieraus und aus dem Sachverhalt, dass die X-Variable und die Residualvariable unkorreliert sind, folgt die Zerlegung der Abweichungsquadrate (Quadratsummenzerlegung).

Definition

Bei der **Quadratsummenzerlegung** ergibt sich die Quadratsumme (Summe der quadrierten Abweichungen) einer Variablen Y additiv aus der Quadratsumme von E (d. h. der Summe der quadrierten Regressionsresiduen) und der Quadratsumme von $\hat{Y}$:

$$\sum_{m=1}^{n} (y_m - \bar{y})^2 = \sum_{m=1}^{n} (y_m - \hat{y}_m)^2 + \sum_{m=1}^{n} (\hat{y}_m - \bar{y})^2$$

$$QS_Y = QS_E + QS_{\hat{Y}} \qquad \text{(F 17.16)}$$

Teilt man beide Seiten von Gleichung F 17.16 durch die Anzahl n der Merkmalsträger, erhält man folgende Zerlegung der Varianz der beobachteten y-Werte.

Definition

Bei der **Varianzzerlegung** ergibt sich die Varianz einer Variablen Y additiv aus der Varianz von E und der Varianz von $\hat{Y}$:

$$\frac{\sum_{m=1}^{n} (y_m - \bar{y})^2}{n} = \frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n} + \frac{\sum_{m=1}^{n} (\hat{y}_m - \bar{y})^2}{n}$$

$$s_Y^2 = s_E^2 + s_{\hat{Y}}^2 \qquad \text{(F 17.17)}$$

Systematische und unsystematische Varianz

Die Varianz der y-Werte ist die Summe zweier Varianzkomponenten: der systematischen Varianz $s_{\hat{Y}}^2$, die durch den Prädiktor X gebunden (erklärt, determiniert) wird, und der unsystematischen Varianz s_E^2 (Fehlervarianz, Residualvarianz). Die Unterschiede zwischen den Merkmalsträgern auf dem Merkmal Y lassen sich zum Teil auf ihre Unterschiede auf dem Merkmal X zurückführen, zum Teil aber auch nicht. Ein Teil der Unterschiede in Y ist also nicht auf Unterschiede in X zurückzuführen, sondern auf alle möglichen anderen Einflüsse auf Y, die jedoch nicht mit erhoben wurden. So hängen in unserem Beispiel die Klausurpunkte nicht nur von der Vorbereitungszeit ab, sondern auch von anderen systematischen Unterschieden zwischen den Merkmalsträgern (wie der Intelligenz, dem mathematischen Verständnis oder der Leistungsmotivation) oder unsystematischen Unterschieden (wie etwa den Fehlern, die der Dozent oder die Dozentin bei der Korrektur der Klausur macht). Unsystematische Einflüsse auf Y werden auch als Messfehler bezeichnet (s. Kap. 23). Grundsätzlich gilt: Je größer die Varianz $s_{\hat{Y}}^2$ im Vergleich zur Varianz s_E^2 ist, umso genauer gelingt die Prognose.

17.5 Determinationskoeffizient und Indeterminationskoeffizient

Aus der letztgenannten Feststellung lässt sich ein standardisiertes Maß für die Güte der Vorhersage konstruieren: Da $s^2_{\hat{Y}}$ und s^2_E additiv sind, ist der Anteil von $s^2_{\hat{Y}}$ an der Gesamtvarianz von Y ein Maß dafür, wie präzise die Vorhersage von Y durch X erfolgt. Da $s^2_{\hat{Y}}$ niemals größer sein kann als s^2_Y, kann ein solcher Quotient nur zwischen den Werten 0 und 1 variieren. Ein Wert von 0 würde bedeuten, dass die Varianz der vorhergesagten Werte $s^2_{\hat{Y}} = 0$ wäre und die Gesamtvarianz von Y lediglich auf Fehler zurückzuführen wäre ($s^2_Y = s^2_E$). Ein Wert von 1 würde bedeuten, dass die Varianz der vorhergesagten Werte der Gesamtvarianz entsprechen würde ($s^2_{\hat{Y}} = s^2_Y$); in diesem Fall gäbe es also überhaupt keinen Vorhersagefehler, und die Residualvarianz wäre $s^2_E = 0$.

Der Quotient aus $s^2_{\hat{Y}}$ und s^2_Y, also der Anteil der erklärten Varianz an der Gesamtvarianz, wird als Determinationskoeffizient R^2 bezeichnet:

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y} \qquad \text{(F 17.18a)}$$

Sein Gegenpart, der Quotient aus s^2_E und s^2_Y, also der Anteil der unerklärten Varianz an der Gesamtvarianz, wird als Indeterminationskoeffizient bezeichnet:

$$1 - R^2 = \frac{s^2_E}{s^2_Y} \qquad \text{(F 17.18b)}$$

Determinationskoeffizient und Indeterminationskoeffizient ergeben also in der Summe 1:

$$R^2 + \left(1 - R^2\right) = \frac{s^2_{\hat{Y}}}{s^2_Y} + \frac{s^2_E}{s^2_Y} = 1 \qquad \text{(F 17.19)}$$

Man greift bei der Bezeichnung der beiden Koeffizienten auf R^2 zurück, da beide Koeffizienten eng mit der Produkt-Moment-Korrelation r zusammenhängen. Wie man durch Umformen von Formel F 17.14 leicht erkennen kann (s. Übung 3), gilt im Falle der einfachen linearen Regression:

$$1 - R^2 = 1 - r^2_{XY} \qquad \text{(F 17.20a)}$$

$$R^2 = r^2_{XY} \qquad \text{(F 17.20b)}$$

Der Determinationskoeffizient wird mit einem großen R^2 und nicht mit einem kleinen r^2 bezeichnet, da er auch für den Fall mehrerer unabhängiger Variablen definiert wird (s. Abschn. 19.6) und im Fall mehrerer Variablen das Quadrat der multiplen Korrelation darstellt.

Determinationskoeffizient gleich 0 ($R^2 = 0$)

Ist der Determinationskoeffizient gleich 0, so bedeutet dies, dass beide Variablen unkorreliert sind. Die Variable X ist nicht in der Lage, Unterschiede in Y zu erklären oder vorherzusagen, wenn man einen linearen Zusammenhang voraussetzt. Wie man sich an Gleichung F 17.4 leicht vor Augen führen kann, hat die Regressionsgerade in diesem Fall eine Steigung von 0, denn wenn $r_{XY} = 0$ ist, muss auch $b_1 = 0$ sein:

$$b_1 = 0 \cdot \frac{s_Y}{s_X} = 0$$

Die Regressionsgerade ist dann eine Parallele zur Abszisse. Wo schneidet sie nun die y-Achse, d. h., wo liegt ihr Achsenabschnitt b_0? Wie man sich an Gleichung F 17.5 leicht klarmachen kann, liegt der Achsenabschnitt in diesem Fall bei $b_0 = \bar{y}$:

$$b_0 = \bar{y} - 0 \cdot \bar{x} = \bar{y}$$

Das leuchtet ein, denn wenn X nicht mit Y korreliert ist und insofern keinen Beitrag zur Vorhersage von y-Werten leisten kann, ist der beste Schätzer (d. h. derjenige Wert mit dem geringsten durchschnittlichen Fehler) der unbedingte Mittelwert von Y.

Der Determinationskoeffizient ist immer 0, wenn $\hat{Y}$ eine Konstante ist, da in diesem Fall die Variation in Y nicht auf die Variation in X zurückgeführt werden kann. Ist Y eine Konstante, so ist der Determinationskoeffizient nicht definiert. In diesem Fall wäre es auch nicht sinnvoll, die Variation in Y erklären zu wollen.

Determinationskoeffizient gleich 1 ($R^2 = 1$)

Ist der Determinationskoeffizient gleich 1, bedeutet dies, dass beide Variablen perfekt korreliert sind. Alle Unterschiede in Y lassen sich auf Unterschiede in X zurückführen. Y ist eine lineare Funktion von X, alle Residualwerte und somit die Residualvarianz sind 0. In diesem Fall hat die Regressionsgerade folgende Steigung:

$$b_1 = 1 \cdot \frac{s_Y}{s_X} = \frac{s_Y}{s_X}$$

Der Achsenabschnitt lautet dann:

$$b_0 = \bar{y} - 1 \cdot \frac{s_Y}{s_X} \cdot \bar{x} = \bar{y} - \frac{s_Y}{s_X} \cdot \bar{x}$$

Determinationskoeffizient zwischen 0 und 1 ($R^2 = c$)

Nimmt R^2 einen Wert c zwischen 0 und 1 an, so bedeutet dies, dass $c \cdot 100\,\%$ der Varianz in Y auf die Variation in X zurückgeführt werden können, sofern ein linearer

Zusammenhang zwischen diesen beiden Variablen besteht. Ist der Zusammenhang nicht-linearer Natur, gibt der Determinationskoeffizient in der einfachen linearen Regressionsanalyse nicht den gesamten determinierten Varianzanteil wieder, sondern nur denjenigen Anteil, der auf einen linearen Trend zurückgeführt werden kann. Wie eine kurvilineare Beziehung zwischen zwei Variablen modelliert werden kann und in welcher Weise der Determinationskoeffizient in diesem Fall bestimmt werden kann, behandeln wir in Abschnitt 19.10.

Beispiel

Vorbereitungszeit und Klausurergebnis

Die eingeführten Begriffe und ihre Zusammenhänge wollen wir nun anhand unserer fiktiven Untersuchung zur Vorhersage des Klausurerfolgs aus der Vorbereitungszeit veranschaulichen.

Regressionsresiduum. Tabelle 17.1 (in Abschn. 17.1) enthält in der vierten Wertespalte das Regressionsresiduum e_m. Der Residualwert entspricht dem Schätz- oder Vorhersagefehler, der bei dieser Person gemacht wurde. Für Steinmetz z. B. errechnet sich ein Residualwert von $e_{20} = -5$ (s. Tab. 17.1). Die Summe aller Residualwerte beträgt 0. Folglich beträgt auch der Mittelwert aller Residualwerte 0 (vgl. Tab. 17.1, Zeilen »Summe« und »Mittelwert«).

Fehlervarianz. Die fünfte Wertespalte von Tabelle 17.1 enthält das quadrierte Regressionsresiduum e_m^2. Wenn man die quadrierten Residualwerte aller Merkmalsträger aufsummiert, erhält man die Summe der Abweichungsquadrate (vgl. Zeile »Summe«). Sie beträgt in unserem Beispiel 186. Dividiert man diese Spaltensumme durch die Anzahl der Merkmalsträger, erhält man die empirische Residualvarianz oder empirische Fehlervarianz. In unserem Beispiel ergibt sich für die Fehlervarianz ein Wert von $s_E^2 = 7{,}44$ (vgl. Zeile »Varianz« in der vierten Wertespalte).

Varianzadditivität. Um die Varianzadditivität nachzuvollziehen, berechnen wir zusätzlich zur Fehlervarianz die Varianzen der y-Werte ($s_Y^2 = 32{,}88$) und der $\hat{y}$-Werte ($s_{\hat{Y}}^2 = 25{,}44$). Übereinstimmend mit Formel F 17.17 ist die Differenz zwischen beiden Werten identisch mit dem Wert, den wir für die Fehlervarianz errechnet haben ($s_E^2 = 7{,}44$).

Determinationskoeffizient und Indeterminationskoeffizient. Wenn wir die Varianzen der $\hat{y}$-Werte und der Residualwerte durch die Varianz der y-Werte dividieren, erhalten wir den systematischen Varianzanteil (Determinationskoeffizient) und den unsystematischen Varianzanteil (Indeterminationskoeffizient). Der Determinationskoeffizient beträgt $R^2 = 0{,}77$. Den gleichen Wert erhalten wir, wenn wir die Korrelation zwischen X und Y ($r_{XY} = 0{,}88$) quadrieren. Für den Indeterminationskoeffizienten ergibt sich ein Wert von $1 - R^2 = 0{,}23$. Indeterminationskoeffizient und Determinationskoeffizient addieren sich zu 1.

17.6 Negatives Regressionsgewicht und Regressionsrichtung

Bisher haben wir nur den Fall eines positiven Zusammenhangs behandelt. Im Folgenden thematisieren wir, worin sich ein negativer Zusammenhang zeigt.

17.6.1 Negatives Regressionsgewicht

Wenn X und Y negativ korreliert sind, ergibt sich nach Formel F 17.4 auch ein negatives Regressionsgewicht $b_1 < 0$. Ein negatives Regressionsgewicht bedeutet, dass der erwartete $\hat{y}$-Wert um b_1 Einheiten abnimmt, wenn x um eine Einheit zunimmt. Hätten wir in unserem Beispiel statt der Klausurpunkte Noten auf der üblichen Notenskala von 1 bis 6 verwendet, hätte sich eine negative Korrelation zwischen Vorbereitungszeit und Klausurergebnis ergeben. Der Betrag der negativen Korrelation zwischen Vorbereitungszeit und Klausurnote hätte sich vom Betrag der positiven Korrelation zwischen Vorbereitungszeit und Punktzahl ($r_{XY} = 0{,}88$) höchstwahrscheinlich unterschieden, da die Notenskala keine perfekt lineare Funktion der Punkteskala ist und die Produkt-Moment-Korrelation nur gegenüber linearen Transformationen invariant ist (s. Abschn. 16.3.1). Bei Verwendung der Notenskala statt der Punkteskala hätten sich nicht nur das Vorzeichen und der Betrag der Korrelation geändert, sondern auch die Regressionsgleichung. Denn durch die Transformation von Punkten in Noten ändert sich der Maßstab der abhängigen Variablen, damit das Regressionsgewicht b_1, mit dem X

zur optimalen Vorhersage von Y multipliziert wird, und schließlich auch b_0, der Achsenabschnitt. Allerdings setzt die inferenzstatistische Absicherung der Regressionsgewichte voraus, dass die Residualvariable stetig und normalverteilt ist, was bei der Notenskala nicht der Fall wäre. Man würde daher für die Vorhersage der Note ein Regressionsmodell für kategoriale Variablen mit geordneten Antwortkategorien (s. Abschn. 22.10) vorziehen.

17.6.2 Regressionsrichtung

Bei der Berechnung von Korrelationen wird keine Unterscheidung zwischen abhängigen und unabhängigen Variablen vorausgesetzt oder getroffen. Korrelationskoeffizienten sind symmetrische Zusammenhangsmaße. Bei der Regression gilt dieses Symmetrieprinzip grundsätzlich nicht. Es macht also einen Unterschied, welche der beiden Variablen als unabhängige Variable und welche als abhängige Variable betrachtet wird. Diese Asymmetrie kann man sich leicht anhand von Formel F 17.4 klarmachen, nach der sich das Regressionsgewicht von X (UV) für die Vorhersage von Y (AV) aus der Produkt-Moment-Korrelation und den Standardabweichungen der beiden Variablen berechnen lässt.

Regression von *X* auf *Y*

Wenn wir den Status der beiden Variablen vertauschen und X aus Y vorhersagen, so dass die Regressionsgleichung

$$\hat{x}_m = b_0^* + b_1^* \cdot y_m \qquad \text{(F 17.21)}$$

lautet, ändert sich Formel 17.4 zu

$$b_1^* = r_{XY} \cdot \frac{s_X}{s_Y}, \qquad \text{(F 17.22)}$$

und entsprechend ändert sich auch Formel 17.5 zu

$$b_0^* = \bar{x} - b_1^* \cdot \bar{y}. \qquad \text{(F 17.23)}$$

Für unser Beispiel »Vorbereitungszeit und Klausurergebnis« ergeben sich für die Vorhersage der Klausurpunkte (Y) aus der Vorbereitungszeit (X) und der Vorbereitungszeit (X) aus den Klausurpunkten (Y) folglich zwei verschiedene Regressionsgleichungen:

$$\hat{y}_m = 10 + 0{,}5 \cdot x_m \qquad \text{(F 17.24)}$$

und

$$\hat{x}_m = -8{,}69 + 1{,}55 \cdot y_m \qquad \text{(F 17.25)}$$

Während wir anhand von Gleichung F 17.24 vorhersagen, wie viele Punkte eine Person voraussichtlich erzielt, die sich eine bestimmte Zeit lang auf die Klausur vorbereitet hat, erlaubt Gleichung F 17.25 die Vorhersage (oder genauer: die »Nachhersage«) der Vorbereitungszeit aus der Anzahl der Klausurpunkte.

Regressionsgerade. Anhand eines einfachen Zahlenbeispiels kann man sich davon überzeugen, dass die beiden Regressionsgeraden nicht deckungsgleich durch den Punkteschwarm gehen: Wenn wir vorhersagen, wie viele Punkte jemand voraussichtlich erreicht, der sich $x_m = 10$ Stunden auf die Klausur vorbereitet hat, ergibt sich nach der Regressionsgleichung F 17.24 ein Vorhersagewert von $\hat{y}_m = 15$ Klausurpunkten. Wenn wir die Regressionsrichtung umkehren und vorhersagen (bzw. »nachhersagen«), wie viele Stunden sich jemand auf die Klausur vorbereitet hat, der $y_m = 15$ Klausurpunkte erzielt hat, ergibt sich nach der Regressionsgleichung F 17.25 ein Vorhersagewert (bzw. »Nachhersagewert«) von $\hat{x}_m = 14{,}56$ Stunden.

Abbildung 17.2 illustriert die Asymmetrie der Regression grafisch: Die beiden Regressionslinien verlaufen weder deckungsgleich noch parallel durch den Punkteschwarm, sondern schneiden sich und bilden dabei vier Winkel. Der Cosinus des spitzen Winkels (α) ist mit der Produkt-Moment-Korrelation von X und Y identisch.

Regressionskoeffizienten. Man beachte weiterhin, dass die beiden Regressionsgeraden in Abbildung 17.2 nicht nur unterschiedlich verlaufen, sondern auch unterschiedlich zu lesen sind: Während b_0 den Ordinatenabschnitt bezeichnet, also denjenigen $\hat{y}$-Wert, den wir für einen x-Wert von $x = 0$ erwarten, steht b_0^* für den Abszissenabschnitt, also denjenigen $\hat{x}$-Wert, den wir für einen y-Wert von $y = 0$ vorhersagen. In unserem Beispiel beträgt der Abszissenabschnitt $b_0^* = -8{,}69$. Praktisch ist dieser Wert nicht sinnvoll, da negative Vorbereitungszeiten unvorstellbar sind. Dennoch handelt es sich um den theoretischen Wert, den man bei Personen erwarten würde, die in der Klausur keinen einzigen Punkt erzielt haben.

Auch b_1 und b_1^* haben verschiedene Bedeutung: Das Regressionsgewicht b_1 gibt an, um wie viele Einheiten sich die $\hat{y}$-Werte verändern, wenn wir Y aus X vorhersagen und sich X um eine Einheit verändert. Hingegen gibt das Regressionsgewicht b_1^* an, um wie viele Einheiten sich die $\hat{x}$-Werte verändern, wenn wir X aus Y vorhersagen und sich Y um eine Einheit verändert.

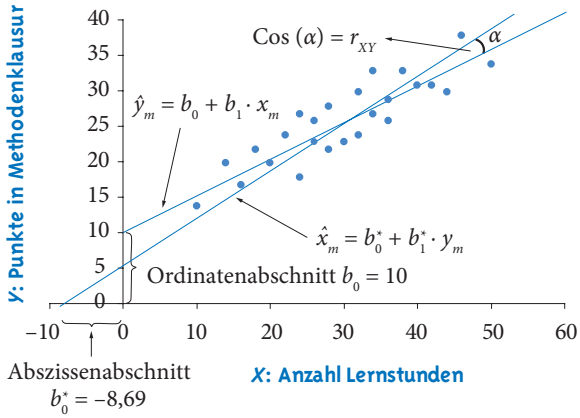

Abbildung 17.2 Regression von *Y* auf *X* und Regression von *X* auf *Y* zum Zahlenbeispiel aus Tabelle 17.1

17.7 Regression standardisierter Werte

Standardisiert man die Werte der unabhängigen und der abhängigen Variablen durch eine *z*-Transformation, dann haben beide Variablen einen Mittelwert von 0 und eine Standardabweichung von 1. Berechnet man auf der Grundlage der standardisierten Werte eine einfache lineare Regression, so ergibt sich folgendes Regressionsgewicht (der Index s zeigt an, dass die Werte zunächst standardisiert wurden):

$$b_{1s} = r_{XY} \cdot \frac{s_Y}{s_X} = r_{XY} \cdot \frac{1}{1} = r_{XY} \qquad \text{(F 17.26)}$$

Bei *z*-standardisierten Variablen ist das Regressionsgewicht mit dem Korrelationskoeffizienten identisch. Für den Ordinatenabschnitt b_{0s} der standardisierten Regression gilt:

$$b_{0s} = \bar{y} - b_{1s} \cdot \bar{x} = 0 - b_{1s} \cdot 0 = 0 \qquad \text{(F 17.27)}$$

Bei einer standardisierten Regression geht die Regressionsgerade also immer durch den Ursprung des Koordinatensystems (0; 0). Die Steigung ist immer gleich der Korrelation. Je steiler die Gerade, desto höher die Korrelation, wobei die Steigung nicht größer als 1 oder kleiner als −1 werden kann. Das Regressionsgewicht der *un*standardisierten Regression hat hingegen keine Ober- oder Untergrenze. Wird die Steigung der Geraden in der standardisierten Regression steiler, wandern die Punkte auch näher zur Geraden hin, da bei höherer Korrelation auch der Punkteschwarm immer enger wird und mit steigender Korrelation gegen eine Gerade konvergiert. Dies ist bei der unstandardisierten Regression nicht zwangsläufig der Fall. Im Falle der unstandardisierten Regression kann die Regressionsgerade sehr steil, die Korrelation hingegen dennoch gering sein, wenn die Varianz von *Y* größer ist als die Varianz von *X* (s. Formel F 17.4). In der standardisierten Regression wird das Regressionsgewicht als Beta-Koeffizient (β) bezeichnet. Wir vermeiden dies und nennen es vielmehr b_{1s}, da wir griechische Buchstaben für Populationsparameter (s. Abschn. 17.9) verwenden. Die standardisierte Regression und die unstandardisierte Regression eröffnen unterschiedliche Einsichten in die Daten. Beide haben ihre Berechtigung. Auf welche Regression sollte man in einer empirischen Untersuchung zurückgreifen?

Unstandardisierte Regression

Auf die unstandardisierte Regression greift man immer dann zurück, wenn man *y*-Werte anhand der *x*-Werte voraussagen will und wenn man den Zusammenhang zwischen verschiedenen Gruppen vergleichen will.

Vorhersage. Will man beispielsweise die Länge der Publikationsliste (*Y*-Variable) einer Wissenschaftlerin anhand ihrer eingeworbenen Forschungsmittel (*X*-Variable) vorhersagen, greift man auf die unstandardisierte Regression zurück. Aufgrund der Regressionsgewichte kann man die erwartete Anzahl von Publikationen vorhersagen und erhält einen Wert, den man direkt interpretieren kann, nämlich als Anzahl von Publikationen bzw. Anteilen davon. Diese einfache Interpretation geht durch die Standardisierung verloren. Das Problem der Standardisierung in diesem Kontext ist, dass sie sich immer auf die Verteilungskennwerte der untersuchten Stichprobe (Mittelwerte, Standardabweichungen) bezieht und daher nicht einfach auf andere Stichproben und Situationen übertragbar ist. Die Maßeinheiten in der unstandardisierten Regression (hier: Euro und Publikationsanzahl) hängen nicht wie bei der standardisierten Regression (Maßeinheit: eine Standardabweichung) von den Streuungen der Merkmale in der Stichprobe ab.

Vergleich verschiedener Gruppen. Will man beispielsweise wissen, ob der Zusammenhang zwischen der Länge der Publikationsliste und der Höhe eingeworbener Forschungsmittel in zwei unterschiedlichen wissenschaftlichen Disziplinen (z. B. Psychologie und Biologie) gleich ist, würde man ebenfalls eher unstandardisierte Regres-

sionsgewichte vergleichen. Gleiche b_1-Koeffizienten würden anzeigen, dass sich in beiden Disziplinen mit einem Euro Forschungsmittelzuwachs der erwartete Publikationsertrag in gleicher Weise ändert. Der Vergleich der standardisierten Regressionsgewichte würde einen anderen Vergleich beinhalten: Angenommen, die b_1-Gewichte wären gleich, aber in der Psychologie wäre die Varianz der eingeworbenen Forschungsmittel geringer als in der Biologie. Dann würden sich unterschiedliche Regressionsgewichte b_{1s} ergeben. Diese unterschiedlichen Regressionsgewichte b_{1s} würden nicht erkennen lassen, dass der erwartete Publikationsertrag in beiden Disziplinen in gleicher Weise von den Forschungsmitteln abhängt. Der Vergleich der standardisierten Regressionskoeffizienten ist deswegen jedoch nicht uninteressant: Unterschiede in den standardisierten Regressionsgewichten würden Korrelationsunterschiede anzeigen und somit den Sachverhalt, dass in beiden Disziplinen die Publikationsleistung unterschiedlich genau (präzise) anhand der Forschungsmittel vorhergesagt werden könnte. Dieser Aspekt ist ebenfalls interessant, hätte aber eine vollkommen andere Bedeutung als die Gleichheit der unstandardisierten Regressionsgewichte.

Standardisierte Regressionsgewichte

Auf standardisierte Regressionsgewichte greift man zurück, wenn man in verschiedenen Studien Zusammenhänge zwischen denselben Merkmalen untersucht hat, die Messinstrumente sich jedoch in ihrer Maßeinheit unterscheiden. Hat man beispielsweise in verschiedenen Ländern den Zusammenhang zwischen Extraversion und Wohlbefinden untersucht, in jedem Land aber eine andere Extraversions- und Wohlbefindensskala eingesetzt, dann ist es wenig sinnvoll, die unstandardisierten Regressionsgewichte zu vergleichen, da sich die Skalen in ihren Maßeinheiten unterscheiden. In diesem Fall vergleicht man die standardisierten Regressionsgewichte. Man hat dann eine vergleichbare Maßeinheit und weiß, dass das standardisierte Regressionsgewicht die erwartete Veränderung im Wohlbefinden pro Extraversionszuwachs um eine Standardabweichung widerspiegelt.

Auf standardisierte Regressionskoeffizienten würde man auch zurückgreifen, wenn man die Zusammenhänge zwischen solchen unterschiedlichen Merkmalen vergleichen wollte, die nicht auf derselben Skala gemessen wurden, z. B. den Zusammenhang zwischen Ärgerintensität und Blutdruck mit dem Zusammenhang zwischen Angstintensität und Hautleitfähigkeit.

17.8 Bedeutung der linearen Regression

Die lineare Regressionsanalyse hat in der Statistik eine große Bedeutung, die weit über die Problemstellungen, die wir in diesem Kapitel behandelt haben, hinausgeht. Sie wird in vielen Bereichen der Sozial- und Verhaltenswissenschaften zur *Prädiktion* von Merkmalsausprägungen, aber auch zur *Erklärung* von Merkmalsunterschieden eingesetzt. Die einfache lineare Regressionsanalyse ist jedoch in dreierlei Hinsicht beschränkt.

Beschr□nkungen

Additiv≠lineare Zerlegung. Die additiv-lineare Zerlegung ist nur dann sinnvoll, wenn die abhängige Variable eine metrische Variable ist. Im Falle nicht-metrischer Variablen greift man auf andere Ansätze wie z. B. die logistische Regressionsanalyse zurück (s. Kap. 22).

Eine unabh□ngige Variable. Die einfache lineare Regressionsanalyse nimmt darüber hinaus an, dass es nur eine unabhängige Variable gibt. Aufgrund der Multideterminiertheit des Verhaltens und Erlebens benötigt man zur Prädiktion und Erklärung üblicherweise jedoch mehrere unabhängige Variablen. Die multiple Regressionsanalyse ist eine diesbezügliche Erweiterung der einfachen Regressionsanalyse (s. Kap. 19).

Linearer Zusammenhang. Schließlich nimmt die einfache lineare Regressionsanalyse an, dass der Zusammenhang zwischen beiden Variablen linear ist. Im Falle nichtlinearer Zusammenhänge kann z. B. auf Spezialfälle der multiplen Regressionsanalyse zurückgegriffen werden (s. Abschn. 19.10). Bevor die multiple Regressionsanalyse behandelt wird, müssen zunächst einige wesentliche Grundkonzepte wie die Partial- und die Semipartialkorrelation im nächsten Kapitel behandelt werden.

17.9 Inferenzstatistik der einfachen linearen Regression

Im Folgenden zeigen wir, wie das Populationsmodell der einfachen linearen Regression aussieht, welche Annahmen getroffen werden müssen, um die Populationsparameter anhand von Stichprobendaten zu schätzen, und wie entsprechende Konfidenzintervalle bestimmt werden können. Wir werden sehen, dass

man bei der Regressionsanalyse – genau wie bei der Varianzanalyse (s. Abschn. 13.1.11) – zwei Modelle unterscheidet:

(1) ein Modell, bei dem man annimmt, dass die Werte der unabhängigen Variablen feste Werte sind, die vollständig durch die Untersuchungsplanung determiniert sind (Modell mit deterministischem Regressor oder »fixed *X* regression model«);
(2) ein Modell, bei dem angenommen wird, dass die Ausprägungen der unabhängigen Variablen Realisierungen einer Zufallsvariablen *X* sind, also nicht vollständig durch die Untersuchungsplanung determiniert werden können (Modell mit stochastischem Regressor oder »random *X* regression model«).

Wir werden die Konsequenzen beider Modelle behandeln und sehen, dass man unter bestimmten Umständen in beiden Modellen auf dieselben Formeln zur Berechnung der Prüfgrößen und Konfidenzintervalle zurückgreifen kann.

17.9.1 Populationsmodell der einfachen linearen Regression

Wir haben das Grundprinzip der linearen Regressionsanalyse anhand der bedingten Mittelwerte eingeführt und aufgezeigt, dass ein lineares Regressionsmodell dann ein sinnvolles Modell zur Beschreibung des Zusammenhangs ist, wenn die Mittelwerte von *Y*, die man für verschiedene *x*-Werte erhält, auf einer Geraden liegen. Wir haben aber auch gesehen, dass diese Annahme auf der Ebene der Stichprobe wenig sinnvoll ist, da die bedingten Mittelwerte schon allein zufallsbedingt von der Geraden abweichen können. Bei der Behandlung der Varianzanalyse (s. Kap. 13) haben wir ein ähnliches Phänomen kennengelernt. Auch wenn die Erwartungswerte der Variablen *Y* in den verschiedenen Bedingungen (Ausprägungen der unabhängigen Variable *X*) in der Population gleich sind, können sie sich in den Stichprobendaten allein aufgrund des Stichprobenfehlers unterscheiden.

Bedingte Erwartung

Im Populationsmodell der einfachen linearen Regression trifft man nun genau die Annahme, dass die Erwartungswerte der Variablen *Y* für jede Ausprägung der Variablen *X* auf einer Geraden liegen:

$$E(Y \mid X) = \beta_0 + \beta_1 \cdot X \qquad \text{(F 17.28)}$$

Der Ausdruck $E(Y|X)$ bezeichnet die bedingte Erwartung der Variablen *Y* gegeben die Variable *X*. Die Werte der bedingten Erwartung sind die bedingten Erwartungswerte $E(Y|X = x)$, d. h. die Erwartungswerte von *Y* für spezifische Werte *x* der unabhängigen Variablen *X* (zum Konzept der bedingten Erwartung s. ausführlich Steyer, 2003). Hieraus ergibt sich die Zerlegung der abhängigen Variablen *Y*:

$$Y = E(Y \mid X) + \varepsilon = \beta_0 + \beta_1 \cdot X + \varepsilon \qquad \text{(F 17.29)}$$

Dabei bezeichnet ε die Residualvariable auf der Populationsebene.

Lineare Regression vs. lineare Quasi-Regression

Die lineare Regression setzt voraus, dass alle bedingten Erwartungswerte auf einer Geraden liegen. Das ist eine sehr strenge Form der Abhängigkeit, die jedoch die Realität nicht immer zutreffend beschreibt. Wie wir in Kapitel 19 sehen werden, gibt es verschiedene Formen der regressiven Abhängigkeit der abhängigen Variablen *Y* von einer unabhängigen Variablen *X*. Diese Abhängigkeit kann auch so geartet sein, dass sie nicht durch eine einfache Funktion beschrieben werden kann. Nähert man die Abhängigkeit der bedingten Erwartung von der unabhängigen Variablen durch eine mathematische Funktion an, so beschreibt diese Funktion nicht mehr die Abhängigkeit der bedingten Erwartung von der unabhängigen Variablen *X*, sondern die bestmögliche funktionale Beschreibung der Abhängigkeit. Zur Unterscheidung zwischen der wahren Regression, die voraussetzt, dass die bedingten Erwartungswerte von *Y* exakt modelliert werden, und der bestmöglichen funktionalen Repräsentation des Zusammenhangs verwendet Steyer (2003) die Unterscheidung zwischen Regression und Quasi-Regression. Diese Unterscheidung wollen wir an einem einfachen Beispiel veranschaulichen.

In Abbildung 17.3 ist die Abhängigkeit der bedingten Erwartung $E(Y|X)$ von *X* dargestellt. Man sieht, dass die bedingten Erwartungswerte von *Y* nicht auf einer Linie angeordnet werden können, sondern um eine Gerade streuen. Dies ist ein Beispiel dafür, dass die bedingte Erwartung in nicht-linearer Weise von der unabhängigen Variablen *X* abhängt, die Abhängigkeit durch eine lineare Gerade jedoch approximiert werden kann. Bei der linearen Geraden handelt es sich aber nicht um die Regression, sondern um die bestmögliche lineare

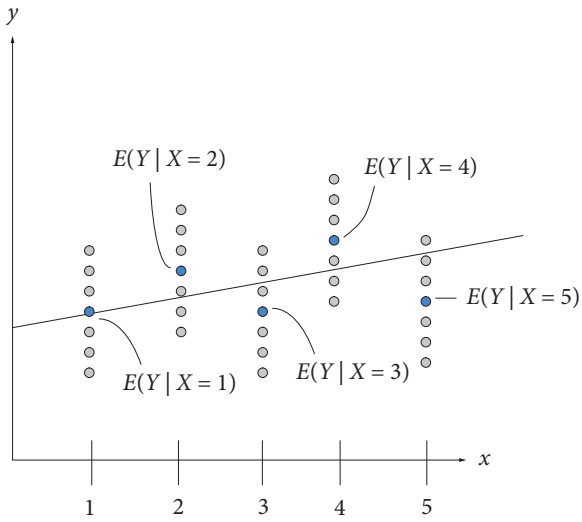

Abbildung 17.3 Nicht-lineare Abhängigkeit der bedingten Erwartung $E(Y|X)$ von X und lineare Quasi-Regression

Repräsentation der Abhängigkeit. Man spricht daher streng genommen von der linearen Quasi-Regression. Eine häufige Anwendung besteht darin, die beste lineare Approximation im Sinne der linearen Quasi-Regression auf der Basis des Kleinste-Quadrate-Kriteriums aufzudecken. Steyer (2003, s. Kap. 8.4) zeigt, wie die Annahme der Linearität der Regression überprüft werden kann. In Kapitel 19 werden wir eine einfache Methode kennenlernen, wie man überprüfen kann, ob eine nicht-lineare Kurve die Abhängigkeit besser beschreibt als eine lineare.

17

17.9.2 Inferenzstatistische Schätzung und Testung

Um die Parameter des Regressionsmodells inferenzstatistisch absichern und die Konfidenzintervalle schätzen zu können, müssen zusätzliche Annahmen getroffen werden (z. B. Fahrmeir et al., 1996b). Wir werden diese Annahmen für das Modell mit einem deterministischen und das Modell mit einem stochastischen Regressor getrennt behandeln.

Modell mit einem deterministischen Regressor

Bei diesem Modell wird angenommen, dass die Ausprägungen der unabhängigen Variablen X feste Werte sind, die durch die Untersuchungsplanung determiniert sind. Man geht von deterministischen unabhängigen Variablen aus, deren Werte vorher ausgewählt werden und sich messfehlerfrei bestimmen lassen. Dies trifft für geplante Untersuchungen zu, in denen man für feste Werte der unabhängigen Variablen Stichproben auswählt, die sich dann bezüglich ihrer y-Werte unterscheiden können. Bezogen auf unser Beispiel wäre dies dann der Fall, wenn in der Untersuchung vorher die Vorbereitungszeiten für die Klausur festgelegt werden würden, man für die ausgewählten Vorbereitungszeiten Substichproben ziehen würde und dann die Werte der abhängigen Variablen (Klausurerfolg) in den einzelnen Substichproben bestimmen würde. Man könnte dann z. B. die Mittelwerte der Punkte in der Methodenklausur bestimmen und eine lineare Regressionsgerade anpassen. Bei der abhängigen Variablen handelt es sich um eine Zufallsvariable, deren Werte man erklären bzw. vorhersagen möchte. Das Modell mit einem deterministischen Regressor wird auch das »klassische« Modell der Regressionsanalyse genannt und ist ein Spezialfall des Allgemeinen Linearen Modells (ALM), das wir in Kapitel 19 ausführlicher behandeln werden. Es entspricht konzeptuell dem varianzanalytischen Modell mit festen Faktoren, das – wie wir in Kapitel 19 sehen werden – ebenfalls einen Spezialfall des ALM darstellt.

Zur inferenzstatistischen Testung der Parameter und Bestimmung der Konfidenzintervalle im Modell mit einem deterministischen Regressor müssen drei zusätzliche Annahmen getroffen werden (Mickey et al., 2004):

(1) Homoskedastizität
(2) Normalverteilung der Fehlervariablen
(3) Unabhängigkeit der Fehler

Dies sind genau die drei Annahmen, die wir bei der Varianzanalyse mit festen Effekten schon kennengelernt haben. In Abschnitt 19.13 werden wir ausführlich behandeln, wie man die Gültigkeit dieser Annahmen überprüfen kann und wie man bei Verletzung der Annahmen verfährt.

Homoskedastitizität. Unter Homoskedastizität versteht man, dass die bedingte Varianz $Var(Y|X)$ in der Population für jede Ausprägung x von X gleich ist. $Var(Y|X = x)$ ist die Varianz der abhängigen Variablen für einen spezifischen x-Wert. Im Modell der einfachen linearen Regression ist es die Streuung (Varianz) der y-Werte um die Regressionsgerade an der Stelle einer bestimmten Merkmalsausprägung x. Die bedingte Varianz $Var(Y|X)$ ist gleich der bedingten Fehlervarianz

$Var(\varepsilon|X)$, denn die Streuung der y-Werte um die Regressionsgerade (d. h. um einen bedingten $\hat{y}$-Wert) ist nichts anderes als die Streuung der bedingten Regressionsresiduen. Die bedingte Fehlervarianz wird auch mit σ_ε^2 bezeichnet:

$$Var(Y|X) = Var(\varepsilon|X) = \sigma_\varepsilon^2 \qquad \text{(F 17.30)}$$

Bedingte Normalverteilung. Die bedingten Verteilungen der abhängigen Variablen Y bzw. der Fehlervariablen ε müssen nicht nur gleiche Varianzen aufweisen, sondern darüber hinaus auch normalverteilt sein.

Unabhängigkeit der Fehler. Die Stichprobenziehung muss so geartet sein, dass die Fehler (d. h. die Regressionsresiduen) zwischen den Merkmalsträgern voneinander unabhängig sind. Nehmen wir die Tagesform als ein Beispiel: Die Tagesform beeinflusst das Klausurergebnis systematisch, aber sie wurde hier nicht untersucht; also verursacht ihr Einfluss unerklärte Varianz in Y. Angenommen, die Klausurwerte stammen aus verschiedenen Klausuren, die zu unterschiedlichen Tageszeiten geschrieben wurden. In diesem Fall würden die Residuen derjenigen Studierenden, die dieselbe Klausur geschrieben haben, höchstwahrscheinlich voneinander abhängen. Die Annahme unabhängiger Fehler wäre verletzt. Die Unabhängigkeit von Fehlern hat also viel mit der Frage gemeinsam, ob es sich um eine echte Zufallsziehung aus der Population gehandelt hat oder aber um eine Ziehung, die systematische Abhängigkeiten zwischen den Merkmalsträgern (genauer gesagt: zwischen Fehlereinflüssen der Merkmalsträger) begünstigte. Die Unabhängigkeit der Fehler wäre im Falle verschiedener Klausuren, die zu unterschiedlichen Tageszeiten geschrieben wurden, verletzt, da man ein mehrstufiges Auswahlverfahren gewählt hat, in dem zunächst Lehrveranstaltungen per Zufall ausgewählt werden und dann wiederum Studierende pro Lehrveranstaltung per Zufall gezogen werden. In diesem Fall muss die hierarchische Struktur der Daten im Rahmen eines hierarchischen linearen Modells (s. Kap. 20) berücksichtigt werden.

Eine Verletzung der Unabhängigkeitsannahme wäre auch dann gegeben, wenn die Bewertung einer Klausur davon abhängen würde, wie die zuvor korrigierte Klausur bewertet wurde. Dieses Beispiel ist nicht aus der Luft gegriffen: Solche systematischen Reihenfolgeeffekte sind gut untersucht und gut belegt: So neigen Korrektoren dazu, eine mittelmäßige Klausur milder zu bewerten, wenn sie zuvor eine sehr schlechte Klausur korrigieren mussten (Kontrasteffekt). Dieses Problem könnte man dadurch in den Griff bekommen, dass jede Klausur von einem anderen Korrektor korrigiert oder aber man das Bewertungsschema objektivieren würde.

Ist die Annahme unabhängiger Fehler verletzt, so hat dies massive Auswirkungen auf die Wahrscheinlichkeit, eine falsche statistische Entscheidung zu treffen. So erhöht sich die Wahrscheinlichkeit eines α-Fehlers selbst bei schwacher Verletzung der Unabhängigkeitsannahme drastisch: Ein statistischer Test produziert dann mit größerer Wahrscheinlichkeit ein signifikantes Ergebnis, und man würde die Nullhypothese ablehnen, obwohl diese Entscheidung falsch ist (Stevens, 2009).

Die Unabhängigkeit der Fehler bedeutet formal, dass die Fehlervariablen ε_m unabhängig voneinander sind. (Zur Erinnerung: Im Stichprobenmodell erhält die Fehlervariable jetzt einen Index m, da jede Person m die Replikation eines Zufallsvorgangs verkörpert; s. Abschn. 9.3.1.) Bezieht man die beiden anderen Annahmen noch mit ein, lassen sich alle drei Annahmen in der Annahme bündeln, dass die Fehlervariablen ε_m unabhängig und identisch normalverteilt sein müssen mit $\varepsilon_m \sim N(0, \sigma_\varepsilon^2)$. Dies impliziert: $Cov(\varepsilon_m, \varepsilon_{m'}) = 0$ für $m \neq m'$.

Modell mit einem stochastischen Regressor

Im Modell mit einer stochastischen unabhängigen Variablen wird davon ausgegangen, dass die Werte x der unabhängigen Variablen X Realisierungen einer Zufallsvariablen sind, deren empirisch beobachtete Merkmalsausprägungen nicht durch die Untersuchungsplanung kontrolliert werden können. In diesem Modell handelt es sich also nicht nur bei der abhängigen, sondern darüber hinaus auch bei der unabhängigen Variablen um eine Zufallsvariable, deren Werte vor der Durchführung der Untersuchung nicht festliegen, sondern ein Ergebnis der Untersuchung sind. Bezogen auf unser Beispiel würde ein stochastischer Regressor vorliegen, wenn wir per Zufall 25 Studierende auswählen und ihre Vorbereitungszeit und ihr jeweiliges Klausurergebnis registrieren würden. Sowohl die x- als auch die y-Werte sind somit Ergebnisse eines Zufallsprozesses. Dies dürfte der typische Anwendungsfall der Regressionsanalyse in der nicht-experimentellen Forschung sein. Da beide Variablen Zufallsvariablen sind, ist es in diesem Fall möglich, nicht nur die Regression von Y auf X zu betrachten, sondern auch die Regression von X auf Y. Da beide Variablen Zufalls-

variablen sind, ist es auch möglich, anhand der Stichprobe die Populationskennwerte der unabhängigen Variablen (wie z. B. ihren Populationsmittelwert und ihre Populationsvarianz) sowie die Populationskorrelation ρ und die dazugehörigen Standardfehler zu schätzen und Hypothesen zu überprüfen. Hierfür ist es notwendig, Verteilungsannahmen zu treffen. Wie bei der Produkt-Moment-Korrelation wird hierzu üblicherweise die Annahme getroffen, dass beide Variablen bivariat normalverteilt sind. Diese Verteilungsannahme hat den Vorteil, dass hieraus ohne zusätzliche Annahmen folgt, dass die bedingte Verteilung von Y gegeben X normal ist und Homoskedastizität gegeben ist. Dies gilt zwangsläufig auch für die bedingte Verteilung von X gegeben Y. Sowohl für die Regression von Y auf X als auch für die Regression von X auf Y kann daher auf die Schätzverfahren, die für den Fall deterministischer Regressoren entwickelt wurden und die genau auf diesen Annahmen basieren, zurückgegriffen werden. Diese Schätzverfahren werden wir im Folgenden behandeln.

17.9.3 Schätzung der Residualvarianz und des Standardschätzfehlers

Wir zeigen zunächst, wie die Residualvarianz geschätzt werden kann, da diese auch den Standardfehlern der anderen Regressionsparameter zugrunde liegt. Zur Schätzung der Residualvarianz σ_ε^2 greift man auf die Residuen zurück, die man in der Stichprobe anhand des Kleinste-Quadrate-Kriteriums wie folgt erhält: $e_m = y_m - \hat{y}_m = y_m - (b_0 + b_1 \cdot x_m)$. Die durch $n - 2$ geteilte Quadratsumme dieser Residuen ist eine erwartungstreue Schätzung der Residualvarianz in der Population:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{m=1}^{n} e_m^2}{n-2} = \frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n-2} \qquad \text{(F 17.31)}$$

Die Quadratsumme wird durch $n - 2$ geteilt, da man durch die Schätzung von b_0 und b_1 zwei Freiheitsgrade (df) verliert. Bei der Schätzung der Populationsvarianz haben wir gesehen, dass wir durch $n - 1$ teilen müssen, da wir durch die Mittelwertsbestimmung $df = 1$ Freiheitsgrad verlieren. In der Regressionsanalyse ist die Fehlerstreuung die Streuung der y-Werte um die geschätzten $\hat{y}$-Werte. Zur Bestimmung der $\hat{y}$-Werte müssen b_0 und b_1 geschätzt werden, wodurch man nicht nur einen, sondern zwei Freiheitsgrade verliert. Die Wurzel aus dieser geschätzten Populationsfehlervarianz ist der geschätzte Standardschätzfehler, der ein Maß dafür ist, wie stark in der Population die beobachteten Y-Werte um die vorhergesagten Werte streuen:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum_{m=1}^{n} e_m^2}{n-2}} = \sqrt{\frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n-2}} \qquad \text{(F 17.32)}$$

17.9.4 Schätzung und Überprüfung des Regressionsgewichts β_1

Das Regressionsgewicht b_1, das wir in einer Stichprobe anhand der Kleinste-Quadrate-Schätzung bestimmen können (s. Formel F 17.4), stellt eine erwartungstreue Schätzung des Regressionsgewichts β_1 in der Population dar. Unter den drei in Abschnitt 17.9.2 genannten Annahmen folgt die Stichprobenkennwerteverteilung des Regressionsgewichts B_1 ebenfalls einer Normalverteilung mit dem Erwartungswert $E(B_1) = \beta_1$ und der Varianz (quadriertem Standardfehler)

$$Var(B_1) = \sigma_{B_1}^2 = \frac{\sigma_\varepsilon^2}{\sum_{m=1}^{n} (x_m - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{n \cdot s_X^2}. \qquad \text{(F 17.33)}$$

Da die Populationsresidualvarianz σ_ε^2 typischerweise nicht bekannt ist, muss diese aus den Daten geschätzt werden, so dass man folgendermaßen den geschätzten quadrierten Standardfehler von B_1 erhält:

$$\hat{\sigma}_{B_1}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{m=1}^{n} (x_m - \bar{x})^2} = \frac{\hat{\sigma}_\varepsilon^2}{n \cdot s_X^2} \qquad \text{(F 17.34)}$$

Da man s_E^2 nach Formel F 17.14 umformulieren kann in $s_E^2 = (1 - r_{XY}^2) \cdot s_Y^2$ und $\hat{\sigma}_\varepsilon^2 = s_E^2 \cdot n/(n-2)$, folgt hieraus:

$$\hat{\sigma}_{B_1} = \sqrt{\frac{(1 - r_{XY}^2) \cdot s_Y^2 \cdot \frac{n}{n-2}}{n \cdot s_X^2}} = \sqrt{\frac{(1 - r_{XY}^2) \cdot s_Y^2}{(n-2) \cdot s_X^2}}$$
$$= \sqrt{\frac{1 - r_{XY}^2}{n-2} \cdot \frac{s_Y^2}{s_X^2}} \qquad \text{(F 17.35)}$$

Zur Überprüfung der Nullhypothese H_0: $\beta_1 = \beta_{10}$ bzw. ihrer gerichteten Varianten kann auf folgende standardisierte Prüfgröße zurückgegriffen werden, die einer t-Verteilung mit $df = n - 2$ Freiheitsgraden folgt:

$$t = \frac{b_1 - \beta_{10}}{\hat{\sigma}_{B_1}} \qquad \text{(F 17.36)}$$

Hiermit kann auch die spezielle Nullhypothese H_0: $\beta_1 = 0$ überprüft werden. Als zweiseitiges $(1 - \alpha)$-Konfidenzintervall erhält man:

$$b_1 \pm t_{\left(1-\frac{\alpha}{2};\, n-2\right)} \cdot \hat{\sigma}_{B_1} \qquad \text{(F 17.37)}$$

Beispiel

Klausurvorbereitung und Klausurerfolg: Regressionsgewicht

Für unser Anwendungsbeispiel erhalten wir nach Tabelle 17.1:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{m=1}^{n} (y_m - \hat{y}_m)^2}{n-2} = \frac{186}{23} = 8{,}087$$

und

$$\hat{\sigma}_{B_1}^2 = \frac{\hat{\sigma}_\varepsilon^2}{n \cdot s_X^2} = \frac{8{,}087}{25 \cdot 101{,}76} = 0{,}0032$$

und somit $\hat{\sigma}_{B_1} = 0{,}057$. Als zweiseitiges 95 %-Konfidenzintervall erhält man $0{,}5 \pm 2{,}069 \cdot 0{,}057 = 0{,}5 \pm 0{,}118$. Da die 0 nicht im Konfidenzintervall von [0,382; 0,618] liegt, ist das Regressionsgewicht bedeutsam von 0 verschieden. Diese Schlussfolgerung erhält man auch anhand des t-Tests:

$$t = \frac{0{,}5 - 0}{0{,}057} = 8{,}772$$

Der Wert ist größer als der kritische t-Wert von $t_{(0{,}975;\, df = 23)} = 2{,}069$. Die Nullhypothese, dass keine regressive Abhängigkeit vorliegt, wird daher auf einem Signifikanzniveau von $\alpha = 0{,}05$ mit einem zweiseitigen Test verworfen.

17.9.5 Schätzung und Überprüfung des Achsenabschnitts β_0

Auch für den Achsenabschnitt β_0 in der Population ist der Achsenabschnitt b_0, den man mittels der Kleinste-Quadrate-Schätzung anhand der Stichprobendaten nach Formel F 17.5 bestimmen kann, ein erwartungstreuer Schätzer, dessen Stichprobenkennwerteverteilung unter den in Abschnitt 17.9.2 getroffenen Annahmen der Normalverteilung folgt mit dem Erwartungswert $E(B_0) = \beta_0$ und der Varianz

$$\sigma_{B_0}^2 = \sigma_\varepsilon^2 \cdot \frac{\sum_{m=1}^{n} x_m^2}{n \cdot \sum_{m=1}^{n} (x_m - \bar{x})^2} = \sigma_\varepsilon^2 \cdot \frac{s_X^2 + \bar{x}^2}{n \cdot s_X^2}$$
$$= \sigma_\varepsilon^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{n \cdot s_X^2} \right). \qquad \text{(F 17.38)}$$

Muss die Populationsresidualvarianz geschätzt werden, was typischerweise der Fall ist, erhält man folgenden geschätzten Standardfehler:

$$\hat{\sigma}_{B_0} = \hat{\sigma}_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n \cdot s_X^2}} \qquad \text{(F 17.39)}$$

Die Nullhypothese H_0: $\beta_0 = \beta_{00}$ bzw. eine ihrer gerichteten Varianten kann mit folgender Prüfgröße getestet werden, die einer t-Verteilung mit $df = n - 2$ Freiheitsgraden folgt:

$$t = \frac{b_0 - \beta_{00}}{\hat{\sigma}_{B_0}} \qquad \text{(F 17.40)}$$

Als zweiseitiges $(1 - \alpha)$-Konfidenzintervall ergibt sich:

$$b_0 \pm t_{\left(1-\frac{\alpha}{2};\, n-2\right)} \cdot \hat{\sigma}_{B_0} \qquad \text{(F 17.41)}$$

Beispiel

Klausurvorbereitung und Klausurerfolg: Achsenabschnitt

Für unser Anwendungsbeispiel erhalten wir nach Tabelle 17.1:

$$\hat{\sigma}_{B_0} = \hat{\sigma}_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n \cdot s_X^2}}$$
$$= 2{,}844 \cdot \sqrt{\frac{1}{25} + \frac{30^2}{25 \cdot 101{,}76}} = 1{,}785$$

Das zweiseitige 95 %-Konfidenzintervall berechnet sich zu $10 \pm 2{,}069 \cdot 1{,}785 = 10 \pm 3{,}693$. Da die 0 nicht im Konfidenzintervall von [6,307; 13,693] liegt, ist der Achsenabschnitt bedeutsam von 0 verschieden. Diese Schlussfolgerung erhält man auch anhand des t-Tests:

$$t = \frac{10 - 0}{1{,}785} = 5{,}602$$

Der Wert ist größer als der kritische t-Wert von $t_{(0{,}975;\, df=23)} = 2{,}069$, und die Nullhypothese muss auf einem Signifikanzniveau von $\alpha = 0{,}05$ mit einem zweiseitigen Test verworfen werden. Im Gegensatz zur statistischen Absicherung des Regressionsgewichts sind der Signifikanztest und die Bestimmung des Konfidenzintervalls für den Achsenabschnitt häufig wenig interessant. In unserem Anwendungsbeispiel bedeutet die Nullhypothese H_0: $\beta_0 = 0$, dass Studierende, die sich nicht auf die Klausur vorbereitet haben, keinen Punkt erhalten.

17.9.6 Schätzung der bedingten Erwartungswerte

Hat man den Achsenabschnitt b_0 und das Regressionsgewicht b_1 bestimmt, kann man die Regressionsgleichung zur Vorhersage eines Wertes der abhängigen Variablen anhand eines Wertes der unabhängigen Variablen heranziehen. Zum einen kann man den bedingten Erwartungswert $E(Y|X = x)$ schätzen, zum anderen einen individuellen Wert y_m prognostizieren. Wir widmen uns zunächst der Schätzung des Erwartungswerts und behandeln im Abschnitt 17.9.7 die Prognose eines individuellen Wertes.

Den bedingten Erwartungswert $E(Y|X = x)$ schätzt man über den vorhergesagten Wert

$$\hat{y} = b_0 + b_1 \cdot x.$$

Der Standardfehler für die Regressionsgerade beträgt:

$$\hat{\sigma}_{\hat{E}(Y|X=x)} = \hat{\sigma}_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{n \cdot s_X^2}} \qquad \text{(F 17.42)}$$

Je weiter ein x-Wert vom Mittelwert der unabhängigen Variablen abweicht, umso größer ist der Standardfehler des bedingten Erwartungswertes und umso unsicherer ist seine Schätzung. Diesen Sachverhalt kann man sich leicht veranschaulichen, wenn man die einfache lineare Regression in Form von Abweichungswerten darstellt. Ein Abweichungswert ist die Differenz des individuellen Wertes von dem Mittelwert der Variablen. Bildet man für beide Variablen die Abweichungswerte, so erhält man folgende Prognosegleichung im einfachen Regressionsmodell:

$$\hat{y} - \bar{y} = b_1 \cdot (x - \bar{x})$$

Angenommen, der Mittelwert der Y-Variablen in einer Stichprobenuntersuchung sei 2 und das geschätzte Regressionsgewicht b_1 sei 1, so erhält man folgende Prädiktionsgleichung für diesen Anwendungsfall:

$$\hat{y} - 2 = 1 \cdot (x - \bar{x})$$

und somit

$$\hat{y} = 2 + 1 \cdot (x - \bar{x})$$

Jedoch sind sowohl der Mittelwert der Y-Variablen als auch das Regressionsgewicht b_1 geschätzte Werte, die sich im Allgemeinen nicht mit den Populationswerten decken. Wir nehmen nun für unser Beispiel an, dass der wahre Populationsmittelwert von Y den Wert $\mu_Y = 2{,}5$ aufweise und der wahre Populationswert des Regressionsgewichts $\beta_1 = 1{,}5$ betrage. Man erhält dann folgende Prädiktionsgleichung auf der Grundlage der Populationswerte:

$$E(Y|X = x) = 2{,}5 + 1{,}5 \cdot [x - E(X)]$$

Man sieht, dass sich die Stichprobenprädiktionsgleichung von der Populationsprädiktionsgleichung unterscheidet. Man sieht auch, dass der Populationsregressionsparameter $\beta_1 = 1{,}5$ durch den Stichprobenparameter $b_1 = 1$ unterschätzt wird, was auf den Stichprobenfehler zurückgeführt werden kann. Je stärker also ein Wert x vom Mittelwert $\bar{x}$ abweicht, umso stärker werden sich die vorhergesagten Werte anhand der Stichprobengleichung und der Populationsgleichung unterscheiden. Beträgt die Abweichung $x - \bar{x} = 1$, ist der geschätzte Kriteriumswert auf der Grundlage des Stichprobenmodells $\hat{y} = 3$, für das Populationsmodell jedoch $E(Y|X) = 4$. Beträgt die Abweichung $x - \bar{x} = 5$, so ist der vorhergesagte Wert anhand des Stichprobenmodells $\hat{y} = 7$, anhand des Populationsmodells jedoch $E(Y|X) = 10$. Dies zeigt: Je stärker der individuelle Prädiktorwert vom Mittelwert der Prädiktorvariablen abweicht, umso größer ist die Verschätzung des bedingten Erwartungswerts, die auf den Stichprobenfehler zurückgeführt werden kann. Aufgrund dieses Sachverhalts verläuft das Konfidenzintervall für die bedingten Erwartungswerte nicht parallel zur Regressionsgeraden, sondern hat eine bikonkave Gestalt (s. Abb. 17.4). Dieses Konfidenzintervall bestimmt sich wie folgt:

$$\hat{y} \pm t_{(1-\frac{\alpha}{2};\, n-2)} \cdot \hat{\sigma}_{\hat{E}(Y|X=x)} \qquad \text{(F 17.43)}$$

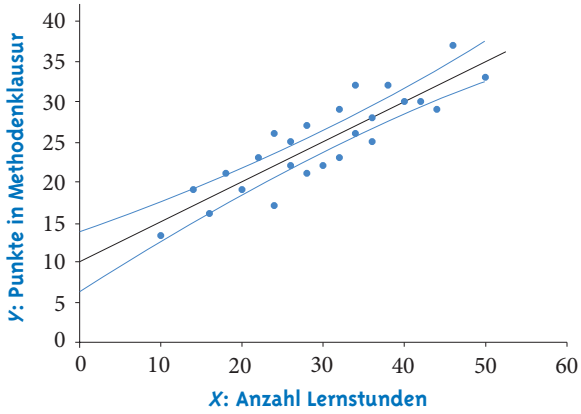

Abbildung 17.4 Konfidenzintervall für die bedingten Erwartungswerte (Regressionsgeraden) zum Zahlenbeispiel aus Tabelle 17.1

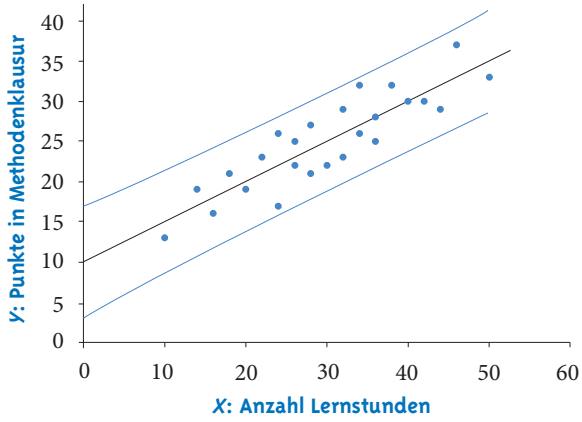

Abbildung 17.5 Konfidenzintervall für vorhergesagte individuelle Werte zum Zahlenbeispiel in Tabelle 17.1

17.9.7 Vorhersage individueller Kriteriumswerte

Will man nicht den bedingten Erwartungswert, sondern einen individuellen y-Wert anhand des x-Wertes vorhersagen, dann ist der Prognosefehler größer. Dies kann man sich leicht veranschaulichen. Will man z. B. für eine Person, die sich 40 Stunden auf die Klausur vorbereitet hat, ihr Klausurergebnis vorhersagen, bevor sie die Klausur geschrieben hat, so wird man als besten Prognosewert den Wert auf der Regressionsgeraden wählen. Zu der Prognoseunsicherheit, die durch die ungenaue Schätzung des bedingten Erwartungswerts zustande kommt (Standardfehler von $\hat{y}$), kommt nun noch eine zweite Quelle der Prognoseunsicherheit ins Spiel, nämlich der Sachverhalt, dass das wirkliche Klausurergebnis nicht gleich dem Wert auf der Regressionsgeraden sein wird, sondern von diesem höchstwahrscheinlich abweichen wird. Bezeichnet man mit x_0 den Wert einer neu gezogenen Person und mit $\hat{y}_0$ den prognostizierten Wert, so erhält man folgende geschätzte Standardabweichung für die geschätzten individuellen Werte:

$$\hat{\sigma}_{\hat{Y}_0} = \hat{\sigma}_\varepsilon \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot s_X^2}} \qquad \text{(F 17.44)}$$

Auch diese Standardabweichung wird umso größer, je stärker der individuelle Wert x_0 vom Mittelwert $\bar{x}$ abweicht. Daher hat auch das Konfidenzintervall

$$\hat{y}_0 \pm t_{(1-\frac{\alpha}{2};\, n-2)} \cdot \hat{\sigma}_{\hat{Y}_0} \qquad \text{(F 17.45)}$$

die typische bikonkave Form, die in Abbildung 17.5 dargestellt ist. Im Vergleich zu Abbildung 17.4 sieht man, dass das Konfidenzintervall für individuelle prognostizierte Werte deutlich größer ist.

17.9.8 Schätzung und Überprüfung des Determinationskoeffizienten

In Abschnitt 16.4.1 haben wir schon gesehen, dass die Produkt-Moment-Korrelation, die man anhand von Stichprobendaten bestimmt, keine erwartungstreue Schätzung der Populationskorrelation ρ darstellt. Entsprechend ist auch die quadrierte Korrelation keine erwartungstreue Schätzung des Determinationskoeffizienten P^2 (P = großes griechisches Rho). Wir haben bereits in Abschnitt 16.4.1 darauf hingewiesen, dass es verschiedene Korrekturformeln gibt, von denen sich die von Olkin und Pratt (1958) vorgestellte Korrektur für verschiedene Anwendungsbereiche als sinnvoll erwiesen hat. Für die Schätzung des Populations-Determinationskoeffizienten lautet die Korrekturformel nach Olkin und Pratt (1958):

$$R^2_{\text{korr OP}} = 1 - \frac{n-3}{n-2} \cdot \left[\left(1 - R^2\right) + \frac{2 \cdot \left(1 - R^2\right)^2}{n} \right] \qquad \text{(F 17.46)}$$

Zur Überprüfung der Nullhypothese H_0: $P^2 = 0$ kann auf folgende Prüfgröße zurückgegriffen werden:

$$F = (n-2) \cdot \frac{R^2}{1-R^2} \qquad \text{(F 17.47)}$$

Diese Prüfgröße folgt unter der Nullhypothese der Unkorreliertheit einer F-Verteilung mit $df_1 = 1$ Zähler- und $df_2 = n - 2$ Nennerfreiheitsgraden. Mit dieser Prüfgröße kommt man zum selben Ergebnis, wie wenn man die Korrelation anhand des t-Tests in Formel F 16.35 zur Signifikanztestung heranzieht. Der F-Wert in F 17.47 ist nichts anderes als der quadrierte t-Wert in F 16.35. Da der Determinationskoeffizient nicht kleiner als 0 werden kann, überprüft man die Nullhypothese H_0: $P^2 = 0$ anhand des kritischen F-Wertes, der an der rechten Seite der Verteilung einen Flächenanteil von α abschneidet.

Ist die Nullhypothese der Unkorreliertheit nicht gültig, so folgt der Ausdruck in Gleichung F 17.47 im Falle eines deterministischen Regressors einer nicht-zentralen F-Verteilung mit dem Nonzentralitätsparameter

$$\lambda = n \cdot \frac{P^2}{1-P^2}. \qquad \text{(F 17.48)}$$

Um ein $(1-\alpha)$-Konfidenzintervall für den Determinationskoeffizienten P^2 zu erhalten, schätzt man zunächst ein Konfidenzintervall für den Nonzentralitätsparameter λ auf der Grundlage des geschätzten Nonzentralitätsparameters

$$\hat{\lambda} = n \cdot \frac{R^2}{1-R^2}. \qquad \text{(F 17.49)}$$

Man transformiert also zunächst das R^2 in den geschätzten Nonzentralitätsparameter, da man für den Nonzentralitätsparameter die Intervallgrenzen einfach bestimmen kann. Hierzu geht man wie folgt vor: Die untere Grenze des Konfidenzintervalls ist der Nonzentralitätsparameter λ_u derjenigen nonzentralen F-Verteilung, von deren oberem Ende der Ausdruck in F 17.47 einen Flächenanteil von $\alpha/2$ abschneidet. Die obere Grenze des Konfidenzintervalls ist der geschätzte Nonzentralitätsparameter λ_o derjenigen nonzentralen F-Verteilung, von deren unterem Ende der Ausdruck in F 17.47 einen Flächenanteil von $\alpha/2$ abschneidet. Diese Intervallschranken können dann durch folgende Gleichungen in die unteren und oberen Grenzen des Intervalls für P^2 umgerechnet werden:

$$P_u^2 = \frac{\lambda_u}{\lambda_u + n} \quad \text{und} \quad P_o^2 = \frac{\lambda_o}{\lambda_o + n} \qquad \text{(F 17.50)}$$

Zur Berechnung der Nonzentralitätsparameter, die die Intervallgrenzen bilden, benötigt man ein Statistikprogramm. Auf geeignete Statistikprogramme wie z. B. das Programm NDC von Steiger (o. J.) verweisen wir in unseren Online-Materialien. Das Konfidenzintervall kann auch direkt mit dem Statistikprogramm R und dem Paket MBESS (Kelley, 2007) berechnet werden.

Im Gegensatz zu Regressionskoeffizienten wird für den Determinationskoeffizienten häufig nicht auf das 95 %-Konfidenzintervall, sondern auf das 90 %-Konfidenzintervall zurückgegriffen. Dies liegt daran, dass der Determinationskoeffizient nur positive Werte annehmen kann. Die Nullhypothese H_0: $P^2 = 0$ wird daher einseitig getestet. Zu einem einseitigen Signifikanztest mit dem Signifikanzniveau α korrespondiert das einseitige $(1-\alpha)$-Konfidenzintervall (s. Abschn. 8.5.2). Es ist aber schwieriger, von einem einseitigen Konfidenzintervall auf die Präzision zu schließen, mit der der Determinationskoeffizient geschätzt wurde (s. Abschn. 8.5.2). Daher präferieren einige Wissenschaftler das zweiseitige Konfidenzintervall (Steiger, 2004). Damit aber das zweiseitige Konfidenzintervall für einen einseitigen Signifikanztest herangezogen werden kann, muss der α-Wert verdoppelt werden. Man kann daher die Nullhypothese, dass P^2 gleich 0 ist, einseitig auf einem a priori festgelegten α-Niveau auch dadurch testen, dass das zweiseitige $(1 - 2 \cdot \alpha)$-Konfidenzintervall berechnet wird und überprüft wird, ob die 0 im zweiseitigen $(1 - 2 \cdot \alpha)$-Konfidenzintervall liegt.

Bestimmung der optimalen Stichprobengröße

Zur Bestimmung der optimalen Stichprobengröße im Falle eines deterministischen Regressors muss man die Effektgröße in der Population sowie das α- und das β-Niveau vorher festlegen. Die Effektgröße lautet:

$$\phi^2 = \frac{P^2}{1-P^2}$$

Nach Cohen (1988) gelten folgende Konventionen:

- kleiner Effekt: $\phi^2 \approx 0{,}02$
- mittlerer Effekt: $\phi^2 \approx 0{,}15$
- großer Effekt: $\phi^2 \approx 0{,}35$

Die optimale Stichprobengröße lässt sich nach Gleichung F 17.48 über den Nonzentralitätsparameter einer nonzentralen F-Verteilung bestimmen. Der Nonzentralitätsparameter – und damit auch die Stichprobengröße – wird so bestimmt, dass der kritische Wert, der unter der zentralen

F-Verteilung (Nullhypothese) einen Flächenanteil von α abschneidet, gleichzeitig unter der nonzentralen F-Verteilung (Alternativhypothese) einen Flächenanteil von $1 - \beta$ nach rechts hin abschneidet. Dieser Nonzentralitätsparameter sowie der dazugehörige optimale Stichprobenumfang kann mit einem Statistikprogramm wie z. B. G*Power (Faul et al., 2007) einfach bestimmt werden.

Unterschiede im Modell mit stochastischem Regressor

Während sich die Stichprobenkennwerteverteilung von R^2 bei Gültigkeit der Nullhypothese H_0: $P^2 = 0$ nicht zwischen dem Modell mit deterministischem und dem Modell mit stochastischem Regressor unterscheidet und man in beiden Fällen dieselben Tests anwenden kann, ist dies nicht mehr der Fall, wenn in der Population ein Zusammenhang zwischen beiden Variablen besteht. Die Stichprobenkennwerteverteilung von R^2 folgt dann nicht mehr einer nonzentralen F-Verteilung. Die Verteilung ist komplexer. Wir wollen diese Verteilung nicht im Detail beschreiben, sondern verweisen auf Mendoza und Stafford (2001). Es gibt jedoch wichtige Konsequenzen sowohl für die Berechnung des Konfidenzintervalls als auch für die Bestimmung der optimalen Stichprobengröße. Beide müssen bei stochastischen Regressoren anders bestimmt werden als bei deterministischen Regressoren. Dies rührt daher, dass man bei stochastischen Regressoren noch den Stichprobenfehler berücksichtigen muss, der mit den unabhängigen Variablen verbunden ist. Tabellen, wie sie z. B. bei Cohen (1988) zu finden sind, gehen üblicherweise von dem Modell mit deterministischem Regressor aus. Für stochastische Regressoren werden spezifische Programme wie z. B. G*Power (Faul et al., 2007) oder R2 (Steiger & Fouladi, 1997) benötigt, die kostenlos im Internet verfügbar sind (Links in den Online-Materialien).

Bei der einfachen linearen Regressionsanalyse kann im Falle eines stochastischen Regressors zur Bestimmung eines Konfidenzintervalls für R^2 auch auf das Konfidenzintervall für die Produkt-Moment-Korrelation, das sich nach Gleichung F 16.45 bestimmen lässt, zurückgegriffen werden. Die Intervallschranken müssen dann quadriert werden. Da der Determinationskoeffizient nur positive, die Korrelation jedoch auch negative Werte annehmen kann, muss die untere Schranke des Konfidenzintervalls für den Determinationskoeffizienten auf 0 gesetzt werden, wenn die 0 ins Konfidenzintervall des Determinationskoeffizienten fällt. Die obere Schranke des Konfidenzintervalls des Determinationskoeffizienten ist dann der größere Wert der beiden quadrierten Schranken des Konfidenzintervalls der Korrelation. Auch bei der Bestimmung der Stichprobengröße kann auf das Verfahren, das wir bei der Produkt-Moment-Korrelation beschrieben haben, zurückgegriffen werden. Dies ist möglich, da im Populationsmodell der Produkt-Moment-Korrelation von zwei stochastischen Variablen ausgegangen wird und die gleiche Verteilungsannahme (bivariate Normalverteilung) wie beim Regressionsmodell mit einem stochastischen Regressor getroffen wird. Während man nach dem in Abschnitt 16.4.1 beschriebenen Vorgehen ein approximatives Konfidenzintervall und eine approximative Stichprobengrößenbestimmung erhält, liefern die Programme G*Power und R2 exakte Werte.

Beispiel

Klausurvorbereitung und Klausurerfolg: Determinationskoeffizient

In unserem Beispiel hat sich eine Korrelation von $r = 0{,}88$ und somit ein Determinationskoeffizient von $R^2 = 0{,}774$ ergeben. Das nach Olkin und Pratt (1958) korrigierte R^2 beträgt $R^2_{\text{korr OP}} = 0{,}780$ und unterscheidet sich nur unwesentlich von dem unkorrigierten Determinationskoeffizienten. Der Wert der Prüfgröße zur Überprüfung der Nullhypothese, dass der Determinationskoeffizient in der Population gleich 0 ist, beträgt $F = 23 \cdot 0{,}774/0{,}226 = 78{,}77$. Berechnet man den F-Wert mit einem Computerprogramm, das mit mehr als drei Nachkommastellen arbeitet, erhält man den präziser geschätzten F-Wert von $F = 78{,}645$. Dieser Wert ist deutlich größer als der kritische F-Wert von $F_{(0{,}95;\,1;\,23)} = 4{,}279$. Die Nullhypothese wird daher verworfen.

Zur Bestimmung des zweiseitigen 90 %-Konfidenzintervalls wurde zunächst angenommen, dass die unabhängige Variable ein deterministischer Regressor sei, was in unserem Beispiel wenig sinnvoll ist, aber der klassischen Annahme des Allgemeinen Linearen Modells entspricht. Zunächst wurde mit dem im Internet frei verfügbaren und sehr einfach zu bedienenden Programm NDC (Steiger, o. J.;) für den gefundenen F-Wert als untere Grenze des Konfidenzintervalls der Nonzentralitätsparameter $\lambda_u = 37{,}330$ und als obere Grenze der Nonzentralitätsparameter $\lambda_o = 132{,}379$ bestimmt. Diese Intervallgrenzen wurden dann in die unteren und oberen Grenzen des

Intervalls für P^2 nach Gleichung F 17.50 umgerechnet, wodurch man $P_u^2 = 0{,}599$ und $P_o^2 = 0{,}841$ und somit das Konfidenzintervall [0,599; 0,841] erhält. Berechnet man das zweiseitige 90 %-Konfidenzintervall für das Modell mit einem stochastischen Regressor mit dem Statistikprogramm R2 (Steiger & Fouladi, 1992), ergibt sich das Konfidenzintervall [0,586; 0,875], das erwartungsgemäß größer ausfällt. Da wir in unserem Beispiel von einem stochastischen Regressor ausgehen, greifen wir auf dieses Konfidenzintervall zurück.

Um die benötigten optimalen Stichprobengrößen zu vergleichen, wurde die optimale Stichprobengröße bestimmt, um einen theoretisch postulierten Determinationskoeffizienten von $P_1^2 = 0{,}25$ bei einem $\alpha = 0{,}05$ und einem $\beta = 0{,}20$ statistisch absichern zu können. Die Berechnung mit dem Statistikprogramm G*Power für das Modell mit einem deterministischen Regressor ergab eine optimale Stichprobengröße von $n = 26$. Die optimale Stichprobengröße für das Modell mit einem stochastischen Regressor, die wir mit R2 berechnet haben, lag erwartungsgemäß mit $n = 29$ höher. Da wir in unserem Beispiel von einem stochastischen Regressor ausgehen, würden wir für unsere Studie eine Stichprobengröße von $n = 29$ festlegen.

Zusammenfassung

- Anhand einer Regressionsanalyse werden Unterschiede in einer abhängigen Variablen Y auf Unterschiede in einer unabhängigen Variablen X zurückgeführt.
- Je nach Forschungsfrage wird die abhängige Variable auch Kriteriumsvariable, zu erklärende Variable oder Regressand genannt und die unabhängige Variable entsprechend Prädiktorvariable, erklärende Variable oder Regressor.
- In einer linearen Regression geht man davon aus, dass in der Population alle bedingten Erwartungswerte der abhängigen Variablen Y, die man für die einzelnen Werte der Variablen X erwartet, auf einer Geraden liegen, der Regressionsgeraden.
- Die Regressionsgerade wird durch den Achsenabschnitt b_0 (Population: β_0) und das Regressionsgewicht b_1 (Population: β_1) bestimmt.
- Der Achsenabschnitt und das Regressionsgewicht lassen sich nach der Kleinste-Quadrate-Methode schätzen.
- Ein Residual- oder Fehlerwert ist die Abweichung des empirisch gefundenen y-Werts von dem aufgrund der Regression erwarteten $\hat{y}$-Wert (dem Wert auf der Regressionsgeraden).
- Die Fehlerstreuung kennzeichnet das Ausmaß der Prognoseunsicherheit und wird auch Standardschätzfehler genannt.
- Der Determinationskoeffizient (Bestimmtheitsmaß) ist der Anteil der Varianz der abhängigen Variablen Y, der durch die Variation der unabhängigen Variablen X determiniert wird.
- Der Determinationskoeffizient ist im Falle der einfachen linearen Regression gleich dem quadrierten Produkt-Moment-Korrelationskoeffizienten.
- Der Determinationskoeffizient kann Werte zwischen 0 (keine Varianzaufklärung) und 1 (100%ige Varianzaufklärung) annehmen.
- Der Indeterminationskoeffizient ist der Anteil der Residualvarianz an der Varianz der abhängigen Variablen Y.
- Auf die unstandardisierten Regressionskoeffizienten greift man zurück, wenn man Gruppen vergleichen oder individuelle Werte vorhersagen will.
- Auf die standardisierten Regressionsgewichte greift man zurück, wenn man Zusammenhänge zwischen Variablen, die in unterschiedlichen Maßeinheiten erfasst wurden, vergleichen will.
- Im Modell mit deterministischem Regressor geht man von festen Werten der unabhängigen Variablen aus, die aufgrund der Untersuchungsplanung feststehen (bzw. kontrolliert werden können) und messfehlerfrei gemessen werden.
- Im Modell mit stochastischem Regressor geht man davon aus, dass die Werte der unabhängigen Variablen Realisierungen einer Zufallsvariablen sind. Die realisierten Werte hängen von der Stichprobenziehung ab.
- Zur inferenzstatistischen Absicherung der Regressionskoeffizienten und zur Bestimmung der Konfidenzintervalle im Modell mit deterministischem Regressor werden die Annahmen der bedingten Normalverteilung, der Homoskedastizität und der Unabhängigkeit der Fehler getroffen.
- Im Modell mit stochastischem Regressor trifft man die Annahme, dass beide Variablen bivariat normalverteilt sind.
- Unter Gültigkeit der Nullhypothese, dass kein Zusammenhang zwischen beiden Variablen besteht,

unterscheiden sich beide Modelle nicht in den statistischen Tests und Konfidenzintervallen.

- Beide Modelle unterscheiden sich in der Verteilung des Determinationskoeffizienten R^2, wenn in der Population ein Zusammenhang zwischen beiden Variablen besteht. Daher muss zur Bestimmung des Konfidenzintervalls des Determinationskoeffizienten und zur Bestimmung der optimalen Stichprobengröße auf unterschiedliche Verfahren zurückgegriffen werden. Im Falle eines stochastischen Regressors ist das Konfidenzintervall größer, und es wird eine größere Stichprobe benötigt.

Fragen und Übungsaufgaben

Fragen

(1) Wie heißt die Bestimmungsgleichung der Regressionsgeraden in der einfachen linearen Regressionsanalyse?

(2) Wie werden der Achsenabschnitt und die Steigung der Geraden in der unstandardisierten und in der standardisierten einfachen linearen Regressionsanalyse bestimmt?

(3) In welche Komponenten kann die Varianz der abhängigen Variablen in der einfachen linearen Regressionsanalyse zerlegt werden?

(4) Wie sind der Determinations- und der Indeterminationskoeffizient definiert, und was bedeuten sie?

(5) Nennen Sie drei Eigenschaften der Residualwerte in der einfachen linearen Regressionsanalyse.

(6) Erläutern Sie das Grundprinzip der Kleinste-Quadrate-Schätzung.

Übungsaufgaben

(1) Tragen Sie die x-Werte und die y-Werte aus Tabelle 17.1 mittels eines Tabellenkalkulationsprogramms oder eines Statistikprogramms in eine Datei ein, und berechnen Sie anhand der Regressionsgleichung F 17.2 die $\hat{y}$-Werte sowie die Residualwerte ($e_m = y_m - \hat{y}_m$). Überprüfen Sie das Ergebnis anhand der Werte in Tabelle 17.1.

(2) Berechnen Sie mithilfe der Korrelationsprozedur des verwendeten Programms die Produkt-Moment-Korrelationen zwischen der unabhängigen Variablen X, der abhängigen Variablen Y, der Variablen $\hat{Y}$ der vorhergesagten Werte und der Residualvariablen E. Vergleichen Sie die Korrelationen, und erklären Sie das Ergebnis.

(3) Zeigen Sie, dass aus Formel F 17.14 und F 17.18b folgt: $1 - R^2 = 1 - r_{XY}^2$ und $R^2 = r_{XY}^2$.