

19 Multiple Regressionsanalyse

Was Sie in diesem Kapitel lernen

- ▶ Brauchen wir mehr als eine unabhängige Variable zur Vorhersage des Wohlbefindens?
- ▶ Wie nützlich sind Persönlichkeitsvariablen zur Vorhersage des Wohlbefindens?
- ▶ Welche Menschen profitieren mehr von ihrer Fähigkeit zur emotionalen Ansteckung als andere Menschen?
- ▶ Hängt die Zufriedenheit mit universitärer Lehre linear von den erlebten Anforderungen ab? Oder sind Studierende umso zufriedener, je weniger sie unter- oder überfordert werden?
- ▶ Was ist Lords Paradox?
- ▶ Welche Voraussetzungen müssen erfüllt sein, damit man die multiple Regressionsanalyse anwenden darf?
- ▶ Wie kann man überprüfen, ob diese Voraussetzungen erfüllt sind?

19.1 Zielsetzungen der multiplen Regressionsanalyse

Mit der Partial- und der Semipartialkorrelation haben wir einfache Methoden zur Kontrolle von Drittvariablen kennengelernt. Mit den Ideen der Partial- und Semipartialkorrelation eng verwandt ist die multiple Regressionsanalyse. Sie stellt eine Erweiterung der einfachen Regressionsanalyse dar, die bereits in Kapitel 17 behandelt wurde. Während in der einfachen Regressionsanalyse Unterschiede in einer abhängigen Variablen (Kriterium, Regressand) auf Unterschiede in einer unabhängigen Variablen (Prädiktor, Regressor) zurückgeführt werden, werden in der multiplen Regressionsanalyse mehrere unabhängige Variablen berücksichtigt. Daher eignet sich die multiple Regressionsanalyse zur Untersuchung der Multideterminiertheit des Verhaltens.

19.1.1 Berücksichtigung von Redundanzen und Kontrolle von Störvariablen

Die multiple Regression berücksichtigt korrelierte Prädiktoren. Korrelierte Prädiktoren können einerseits redundante Information anzeigen, andererseits aber auch darauf hinweisen, dass die regressive Beziehung zwischen zwei Variablen von einer Drittvariablen verfälscht wird. Eine wichtige Zielsetzung der multiplen Regressionsanalyse besteht darin, solche Redundanzen zu erkennen und ggf. zu eliminieren sowie den Einfluss von Störvariablen zu kontrollieren. Diesbezüglich sind die Ziele der multiplen Regressionsanalyse mit den Zielen der Partial- und Semipartialkorrelation vergleichbar. Wie wir im letzten Kapitel gesehen haben, basieren die Partial- und Semipartialkorrelation zweiter und höherer Ordnung auf der multiplen Regressionsanalyse, da sie Korrelationen zwischen Regressionsresiduen darstellen, die bezüglich mehrerer unabhängiger Variablen gebildet wurden. Zwischen der Partial- und Semipartialkorrelation einerseits und der multiplen Regressionsanalyse andererseits besteht jedoch ein wesentlicher Unterschied: Während die Partial- und Semipartialkorrelation einen ungerichteten Zusammenhang zwischen zwei Variablen kennzeichnen, beschreibt die multiple Regression einen gerichteten Zusammenhang. Diese Gerichtetheit ist die Grundlage für zwei weitere Zielsetzungen der multiplen Regressionsanalyse: Prognose (Prädiktion) und Erklärung.

19.1.2 Prognose und Erklärung

Die multiple Regressionsanalyse ermöglicht es, ein Kriterium anhand mehrerer Prädiktoren vorherzusagen. So erlaubt die multiple Regressionsanalyse, den Wert einer Person auf einer Kriteriumsvariablen zu prognostizieren, wenn man die Ausprägungen dieser Person auf den Prädiktorvariablen kennt. Ein Beispiel hierfür ist die Prognose des zukünftigen Berufserfolgs anhand von Prädiktoren, die im Rahmen einer Eignungsuntersuchung erhoben werden. In den meisten Fällen besteht das Ziel der multiplen Regressionsanalyse jedoch in der

Erklärung individueller Unterschiede in der Ausprägung einer Kriteriumsvariablen. Die Frage z. B., warum sich Menschen in Persönlichkeitsvariablen wie der Extraversion oder dem Neurotizismus unterscheiden, lässt sich nur anhand einer Vielzahl potenziell erklärender Variablen untersuchen. So sind neben genetischen Einflüssen auch Umwelteinflüsse zu berücksichtigen, die die Persönlichkeit eines Individuums beeinflussen können. Während sich die Prognose auf die Vorhersage eines zukünftigen (d. h. noch nicht erhobenen) Merkmals bezieht, setzt die Erklärung an vorhandenen (d. h. bereits beobachteten) Merkmalsausprägungen an und will die Bedingungen für deren Unterschiedlichkeit untersuchen.

Beschreiben vs. Erklären von Abhängigkeit. Die Prognose setzt nicht voraus, dass die Prädiktoren auch die Ursachen der zukünftigen Merkmalsausprägung sind. So kann man aufgrund des Extraversionswertes einer Person ihr zukünftiges Wohlbefinden prognostizieren, ohne genau zu wissen, aufgrund welcher Wirkungszusammenhänge die Extraversion das Wohlbefinden beeinflusst. Für die Prognose einer Merkmalsausprägung genügt es also, die regressive Abhängigkeit zwischen der Kriteriumsvariablen und den Prädiktorvariablen lediglich zu beschreiben. Eine Erklärung dieser Abhängigkeiten ist hingegen gehaltvoller, da sie den zugrunde liegenden Wirkmechanismus für den Zusammenhang zwischen Variablen untersucht. So können wir z. B. das Wohlbefinden eines Zwillingsschwagers anhand des Wohlbefindens seines Zwillingsschwagers besser prognostizieren, als wenn wir das Wohlbefinden des Schwagers nicht kennen (z. B. Eid et al., 2003c). Die Erklärung dieses Zusammenhangs ist jedoch schwieriger. Der Zusammenhang kann möglicherweise darauf zurückgeführt werden, dass die Schwäger füreinander sorgen und sich aufmuntern. Es ist jedoch auch denkbar, dass die Zwillinge in ihrer Herkunftsfamilie in ähnlicher Weise gelernt haben, ihre Gefühle zu regulieren. Schließlich könnte der Zusammenhang auch dadurch bedingt sein, dass das Wohlbefinden zu einem gewissen Grad erblich ist.

Bedingungen für Kausalität. Um zwischen verschiedenen Erklärungsmodellen unterscheiden zu können, benötigen wir weitere Informationen. Um die gerichtete Abhängigkeit zwischen zwei Variablen derart kausal interpretieren zu können, dass die unabhän-

gige Variable die Ursache für die abhängige Variable ist, müssen mehrere Voraussetzungen erfüllt sein. So muss die unabhängige Variable der abhängigen Variablen zeitlich vorgeordnet sein, die Ursache also *vor* der Wirkung existieren. Auch darf der regressive Zusammenhang zwischen zwei Variablen nicht auf andere Erklärungsmöglichkeiten, z. B. ausgelassene Drittvariablen, zurückführbar sein. Steyer (1992, 2003) hat auf der Grundlage der Regressionsanalyse eine sehr elaborierte Kausalitätstheorie entwickelt. Starke Kausalität liegt seiner Theorie zufolge vor, wenn sich der Einfluss der unabhängigen Variablen auf die abhängige Variable nicht ändert, wenn weitere unabhängige Variablen in die Regressionsanalyse mit aufgenommen werden. Neben dieser starken Form der Kausalität lassen sich schwächere Formen definieren (s. hierzu ausführlich Steyer, 1992). Einen Überblick über kausalitätstheoretische Ansätze gibt Pearl (2009).

19.1.3 Analyse komplexer Zusammenhänge

Wie wir im Folgenden sehen werden, können Prognose- und Erklärungsmodelle sehr komplex sein. Nicht immer hängt ein Kriterium nur in linearer Weise von einem Prädiktor ab. Vielmehr können zwischen den Prädiktoren und dem Kriterium auch kurvenförmige Zusammenhänge bestehen. Solche nicht-linearen Zusammenhänge werden von manchen Theorien explizit vorhergesagt. Auch ist der Einfluss einer unabhängigen Variablen auf eine abhängige Variable nicht immer für alle Personen gleich. Vielmehr kann die regressive Abhängigkeit zweier Variablen interindividuell variieren und von der Ausprägung einer dritten Variablen abhängen, einer sog. Moderatorvariablen (moderierte Regression).

In der einfachen linearen Regression in Kapitel 17 haben wir nur den Fall untersucht, dass sowohl die abhängige wie auch die unabhängige Variable metrischer Natur sind. Diese Annahme werden wir in Bezug auf die abhängige Variable aufrechterhalten. Die unabhängige Variable hingegen kann metrisch, aber auch dichotom (d. h. kategorial mit zwei Ausprägungen) sein. Wie wir sehen werden, ist die Varianzanalyse ein Spezialfall der Regressionsanalyse mit nominalskalierten unabhängigen Variablen. Da sich die Varianzanalyse und die Regressionsanalyse unter einem gemeinsamen Dach vereinen lassen, nennt man den methodo-

logischen Ansatz, bei dem Merkmalsunterschiede in einer abhängigen metrischen Variablen vorausgesagt bzw. erklärt werden sollen, auch Allgemeines Lineares Modell (ALM). Die Regressionsanalyse mit metrischen unabhängigen Variablen und die Varianzanalyse mit nominalskalierten unabhängigen Variablen sind zwei Spezialfälle des ALM. Im Rahmen des ALM können auch unabhängige Variablen analysiert werden, die sich in ihrem Skalenniveau unterscheiden. So kann z. B. ein Teil der unabhängigen Variablen nominalskaliert und ein Teil der unabhängigen Variablen metrischer Natur sein. Sofern die nominalskalierten Variablen nicht mit den metrischen Variablen interagieren, nennt man diesen Spezialfall Kovarianzanalyse.

Im Folgenden werden wir die Grundprinzipien der multiplen Regressionsanalyse erläutern. Darüber hinaus werden wir zeigen, wie sich komplexere Modelle wie die Kovarianzanalyse und die moderierte Regressionsanalyse anwenden lassen. Schließlich behandeln wir nominalskalierte unabhängige Variablen und Modelle mit unabhängigen Variablen unterschiedlicher Skalenniveaus.

19.2 Notation

Wie bei der einfachen Regressionsanalyse lassen sich verschiedene Notationen unterscheiden. Wir beschreiben zunächst die multiple Regressionsanalyse als ein Modell zur Beschreibung empirisch erhobener Daten. Diese deskriptivstatistische Behandlung der multiplen Regressionsanalyse kommt ohne inferenzstatistische Kenntnisse aus. In Abschnitt 19.7 beschreiben wir dann das Populationsmodell der multiplen Regressionsanalyse und behandeln Fragen der Inferenzstatistik.

Regressionsgleichung für Merkmalsträger

In der Modellgleichung für Merkmalsträger werden die Werte der abhängigen Variablen wie in der einfachen Regressionsanalyse mit y_m bezeichnet. Der Index m steht für den Merkmalsträger (Untersuchungseinheit, z. B. eine Person). In der multiplen Regressionsanalyse werden k verschiedene unabhängige Variablen X_1, \dots, X_k berücksichtigt. Der Wert eines Merkmalsträgers (z. B. einer Person) auf einer unabhängigen Variablen erhält zwei Indizes (x_{mj}): Der erste Index m bezeichnet den Merkmalsträger (z. B. die Person) und läuft von 1 bis n (n : Anzahl der betrachteten Personen); der zweite Index j gibt an, um welche Variable X es sich han-

delt, und läuft von 1 bis k (k : Anzahl der unabhängigen Variablen). Mit e_m wird der Residualwert einer Person m bezeichnet. Hieraus ergibt sich folgende Modellgleichung für Merkmalsträger (Untersuchungseinheiten):

$$y_m = b_0 + b_1 \cdot x_{m1} + b_2 \cdot x_{m2} + \dots + b_j \cdot x_{mj} + \dots + b_k \cdot x_{mk} + e_m, \quad (\text{F } 19.1)$$

wobei mit b_0 der Achsenabschnitt bezeichnet wird und die Koeffizienten b_1 bis b_k die Regressionsgewichte der einzelnen unabhängigen Variablen sind.

Regressionsgleichung für Variablen

Die multiple Regressionsanalyse kann auch in folgender Modellgleichung für Variablen formuliert werden:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_j \cdot X_j + \dots + b_k \cdot X_k + E \quad (\text{F } 19.2)$$

Mit Y wird die abhängige Variable (Kriterium) bezeichnet, während X_1 bis X_k die unabhängigen Variablen kennzeichnen. E bezeichnet die Residualvariable. Die Merkmalsausprägungen y_m sind die Werte der Variablen Y , die Merkmalsausprägungen x_{m1}, \dots, x_{mk} die Werte der Variablen X_1 bis X_k . Die Residualwerte e_m sind die Werte der Variablen E .

Regressionsgleichung für vorhergesagte Werte

Wie bei der einfachen Regressionsanalyse lässt sich für jede gefundene Konstellation der Werte der unabhängigen Variablen ein Wert auf der abhängigen Variablen vorhersagen. Greift man auf die Modellgleichung für Merkmalsträger zurück, so erhält man:

$$\hat{y}_m = b_0 + b_1 \cdot x_{m1} + b_2 \cdot x_{m2} + \dots + b_j \cdot x_{mj} + \dots + b_k \cdot x_{mk} \quad (\text{F } 19.3)$$

Der Wert \hat{y}_m ist der anhand der Regressionsanalyse vorhergesagte Wert einer Untersuchungseinheit (z. B. einer Person) m . Er entspricht somit einer Linearkombination der x -Werte. Die Regressionsgewichte werden so bestimmt, dass der \hat{y}_m -Wert optimal vorhergesagt wird. Was »optimal« bedeutet, werden wir in Abschnitt 19.3.3 behandeln. In Variablenform lässt sich Gleichung F 19.3 wie folgt schreiben:

$$\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_j \cdot X_j + \dots + b_k \cdot X_k \quad (\text{F } 19.4)$$

19.3 Lineare Regression für zwei metrische unabhängige Variablen

Im Folgenden sollen einige Eigenschaften der multiplen Regressionsanalyse für den einfachsten Fall zweier metrischer unabhängiger Variablen erläutert werden. In diesem Fall lässt sich ein beobachteter y -Wert wie folgt zerlegen:

$$y_m = b_0 + b_1 \cdot x_{m1} + b_2 \cdot x_{m2} + e_m \quad (\text{F } 19.5)$$

Ein beobachteter y -Wert lässt sich somit additiv zerlegen in eine Regressionskonstante b_0 , den mit b_1 gewichteten Wert x_{m1} , den mit b_2 gewichteten Wert x_{m2} und den Residualwert e_m .

19.3.1 Multiple Regression als kompensatorisches Modell

Da sich Individuen auf den unabhängigen Variablen X_1 und X_2 unterscheiden können, kann derselbe vorhergesagte \hat{y} -Wert durch eine unterschiedliche Kombination von x_1 - und x_2 -Werten zustande kommen. Den kompensatorischen Charakter der multiplen Regressionsanalyse wollen wir anhand eines Beispiels erläutern.

Beispiel: Habituelles Wohlbefinden und Stimmungsregulationskompetenz

In einer Untersuchung wurde überprüft, inwieweit individuelle Unterschiede im habituellen Wohlbefinden auf die Unterschiede in der Stimmungsregulationskompetenz zurückgeführt werden können. In der Stimmungsregulationstheorie unterscheidet man zwei allgemeine Regulationskompetenzen: die Fähigkeit zur Aufrechterhaltung guter Stimmung und die Fähigkeit zur Verbesserung schlechter Stimmung (Lischetzke & Eid, 2003). Personen mit einer hohen Fähigkeit zur Aufrechterhaltung guter Stimmung sind in der Lage, eine gute Stimmung, die sie erleben, lange aufrechtzuerhalten. Personen mit einer hohen Fähigkeit zur Verbesserung schlechter Stimmung gelingt es relativ schnell, aus einer schlechten Stimmung herauszukommen. In einer Untersuchung an 237 Studierenden ließ sich die Abhängigkeit des habituellen Wohlbefindens (WB) von beiden Regulationskompetenzen wie folgt beschreiben:

$$\widehat{WB} = 1,801 + 0,354 \cdot AS + 0,336 \cdot VS,$$

wobei $b_0 = 1,801$, $b_1 = 0,354$, $b_2 = 0,336$. Die Variable AS erfasst die Fähigkeit zur Aufrechterhaltung guter Stimmung, VS die Fähigkeit zur Verbesserung schlechter Stimmung, \widehat{WB} das aufgrund beider Kompetenzen vorhergesagte Wohlbefinden. Diese Regressionsgleichung zeigt, dass sowohl die Fähigkeit zur Aufrechterhaltung guter Stimmung als auch die Fähigkeit zur Verbesserung schlechter Stimmung einen positiven Einfluss auf das Wohlbefinden haben, und zwar derart, dass bei höherer Fähigkeit zur Aufrechterhaltung guter Stimmung und höherer Fähigkeit zur Verbesserung schlechter Stimmung das Wohlbefinden höher ist.

Die beiden Stimmungsregulationskompetenzen erfassen zwei unterschiedliche, aber miteinander zusammenhängende Fähigkeiten. Beide Facetten der Regulationskompetenz sind zu $r = 0,445$ miteinander korreliert. Diese Korrelation zeigt an, dass Personen mit höherer Fähigkeit zur Aufrechterhaltung guter Stimmung tendenziell auch eine höhere Fähigkeit zur Verbesserung schlechter Stimmung haben. Allerdings sind aufgrund der nicht-perfekten Korrelation auch unterschiedliche Kombinationen der beiden Regulationskompetenzen möglich. Der kompensatorische Charakter der multiplen Regressionsanalyse zeigt sich darin, dass ein geringer Wert auf der Variablen »Fähigkeit zur Aufrechterhaltung guter Stimmung« durch einen hohen Wert auf der Variablen »Fähigkeit zur Verbesserung schlechter Stimmung« ausgeglichen werden kann. Da die beiden Regulationskompetenzen einen fast gleich hohen Regressionskoeffizienten aufweisen, hat z. B. eine Person, die auf der Variablen AS einen Wert von 1 und auf der Variablen VS einen Wert von 3 hat, einen fast identischen vorhergesagten Wert wie eine Person, die auf der Variablen AS einen Wert von 3 und auf der Variablen VS einen Wert von 1 hat. Das additiv-lineare Modell in dieser einfachen Form ist nur für Variablenkonstellationen geeignet, bei denen eine rein additive Verknüpfung sinnvoll ist. Wir werden an späterer Stelle zeigen, wie in der multiplen Regressionsanalyse neben der rein additiven Verknüpfung der Variablen auch ihre multiplikative Verknüpfung berücksichtigt werden kann.

19.3.2 Grafische Darstellung

Die multiple Regressionsanalyse lässt sich in Form von Pfaddiagrammen darstellen. Abbildung 19.1 zeigt eine multiple Regressionsanalyse mit zwei unabhängigen Variablen X_1 und X_2 . Diese sind durch Pfeile mit der abhängigen Variablen Y verbunden. Die Pfeile gehen immer von den unabhängigen Variablen aus und haben ihre Pfeilspitze bei der abhängigen Variablen. An die Pfeile werden die Regressionsgewichte geschrieben. Der Pfeil, der von der Residualvariablen E auf Y zeigt, erhält keine Zahl, da die Residualvariable mit keinem Regressionskoeffizienten verknüpft ist und somit immer ein Gewicht von 1 hat. Der Bogen zwischen den beiden unabhängigen Variablen X_1 und X_2 kennzeichnet einen ungerichteten Zusammenhang und somit die Kovarianz s_{X_1, X_2} bzw. die Korrelation r_{X_1, X_2} der beiden unabhängigen Variablen. In der pfadanalytischen Darstellungsform wird der Achsenabschnitt b_0 , der nichts zur Erklärung interindividueller Unterschiede in der Kriteriumsvariablen beiträgt, nicht berücksichtigt.

In Abbildung 19.2 ist das Pfadmodell für unser Stimmungsregulationsbeispiel dargestellt. Die empirischen Ergebnisse einer Regressionsanalyse trägt man wie folgt in ein Pfaddiagramm ein: Ein Kästchen steht für eine Variable. In das Kästchen schreibt man den Namen der Variablen, also in unserem Beispiel WB für Wohlbefinden, AS für die Aufrechterhaltung guter Stimmung und VS für die Verbesserung schlechter Stimmung. Unter das Kästchen, das die Variable bezeichnet, schreibt man die Varianz der Variablen. Demzufolge bezeichnet 0,372 die Varianz des Wohlbefindens, 0,367 die Varianz der Aufrechterhaltung guter Stimmung und 0,365 die Varianz der Verbesserung schlechter Stimmung. Der Bogen zwischen den beiden unabhängigen Variablen kennzeichnet den Zusammenhang zwischen beiden Variablen. An diesen Bogen schreibt man die Kovarianz der beiden Variablen, die $s_{X_1, X_2} = 0,163$ beträgt. Unter die Kovarianz schreibt man in Klammern die Korrelation, die in unserem Beispiel $r_{X_1, X_2} = 0,445$ beträgt. Die Zahlen an den beiden Pfeilen sind die Regressionsgewichte. In unserem Beispiel ist $b_1 = 0,354$ das Regressionsgewicht der Variablen Aufrechterhaltung guter Stimmung und $b_2 = 0,336$ das Regressionsgewicht der Variablen Verbesserung schlechter Stimmung. An den Pfeil, der die Residualvariable (die Fehlervariable) repräsentiert, schreibt man die Varianz der Fehlervariablen, die im vorliegenden Fall $s_E^2 = 0,248$ beträgt.

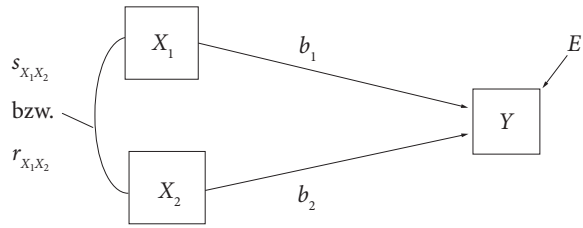


Abbildung 19.1 Pfadanalytische Darstellung einer multiplen Regressionsanalyse mit zwei unabhängigen Variablen. Modellgleichung: $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + E$

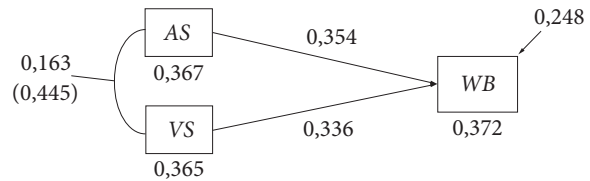


Abbildung 19.2 Ergebnis einer multiplen Regressionsanalyse mit der abhängigen Variablen Wohlbefinden (WB) und den unabhängigen Variablen Fähigkeit zur Aufrechterhaltung guter Stimmung (AS) und Fähigkeit zur Verbesserung schlechter Stimmung (VS)

19.3.3 Bestimmung der Regressionskoeffizienten

Wie bei der einfachen linearen Regression werden auch in der multiplen Regressionsanalyse die Regressionskoeffizienten derart geschätzt, dass das Kleinst-Quadrat-Kriterium erfüllt wird. Dies bedeutet, dass die Summe der Abweichungsquadrate (SAQ) von beobachteten und vorhergesagten y -Werten minimal sein soll.

Definition

Nach dem **Kleinst-Quadrat-Kriterium** wird die Summe der Abweichungsquadrate minimiert:

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

Während bei der einfachen Regressionsanalyse die Regressionskoeffizienten anhand einfacher Formeln berechnet werden können, ist die Bestimmung der Regressionskoeffizienten in der multiplen Regressionsanalyse komplizierter.

Regressionebene

Eine multiple Regressionsanalyse mit zwei unabhängigen Variablen lässt sich auch als Regressionebene darstellen. Diese Ebene entspricht der Regressionsgeraden einer Regressionsanalyse mit einem Prädiktor. In Abbildung 19.3 ist eine multiple Regressionsanalyse mit zwei unabhängigen Variablen X_1 und X_2 dargestellt. Für jede Kombination von x_1 - und x_2 -Werten liegen die geschätzten \hat{y} -Werte auf einer Ebene. Die beobachteten y -Werte schwanken um diese Ebene. Man kann sich das Prinzip wie folgt veranschaulichen. Man stelle sich einen Schwarm Bienen in einem dreidimensionalen Raum vor, z. B. einem Zimmer.

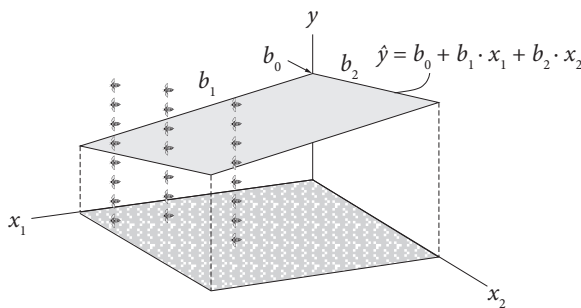


Abbildung 19.3 Darstellung einer multiplen Regressionsanalyse mit zwei unabhängigen Variablen X_1 und X_2 in Form einer Regressionebene

Die Flughöhe einer Biene entspricht dem y -Wert dieser Biene, die Projektion dieser Biene auf den Fußboden entspricht ihren Koordinaten in Bezug auf die Variablen X_1 und X_2 . Durch die Regressionsanalyse wird nun quasi ein großes Blatt durch diesen Bienenschwarm derart gelegt, dass die Summe der quadrierten Abweichungen aller Bienen von diesem Blatt minimal ist.

Regressionsgewichte bei der einfachen und der multiplen Regression

Wie unterscheiden sich die Regressionsgewichte der einzelnen Prädiktoren b_1, \dots, b_k im Falle einer multiplen Regression von den entsprechenden Regressionsgewichten im Falle von k separaten (bivariaten) Regressionsanalysen? Die Regressionsgewichte bei der multiplen Regressionsanalyse sind nur dann mit den entsprechenden Regressionsgewichten der einfachen Regressionsanalyse identisch, wenn die unabhängigen Variablen untereinander unkorreliert sind. Bei korrelierten unabhängigen Variablen muss diese Kor-

relation bei der Berechnung der Regressionsgewichte berücksichtigt werden. Dass dies notwendig ist, kann man sich anhand eines einfachen Beispiels veranschaulichen. Angenommen, eine unabhängige Variable korreliert mit der abhängigen Variablen perfekt. Dann kann aufgrund der unabhängigen Variablen die abhängige Variable genau (d.h. fehlerfrei) vorhergesagt werden. Würde man eine zweite unabhängige Variable, die ebenfalls mit der abhängigen Variablen korreliert ist, in die Regressionsgleichung mit aufnehmen und würde man für b_2 einfach das entsprechende Regressionsgewicht dieses Prädiktors aus der einfachen bivariaten Regression einsetzen, so würde man das Kriterium deutlich verschätzen. Denn das Regressionsgewicht des zweiten Prädiktors würde die Schätzung des Kriteriumswertes verändern – und das, obwohl bereits der erste Prädiktor die Kriteriumswerte perfekt vorhergesagt hat. Bei der Bestimmung der Regressionsgewichte ist es daher wichtig, sowohl die Zusammenhänge zwischen den unabhängigen Variablen untereinander als auch die Zusammenhänge zwischen den unabhängigen Variablen und der abhängigen Variablen zu berücksichtigen.

Die Bestimmung der Regressionsgewichte ist bei mehr als zwei unabhängigen Variablen mathematisch kompliziert und rechnerisch aufwendig; sie wird matrixalgebraisch vorgenommen. Wir werden uns im Folgenden zunächst auf den Fall zweier Prädiktoren konzentrieren, da hierfür die Regressionsgewichte anschaulich bestimmt und interpretiert werden können. Die Bestimmung der Regressionsgewichte bei mehr als zwei Prädiktoren werden wir im Anhang B behandeln.

Bestimmung der Regressionsgewichte b_j

Das Regressionsgewicht der ersten Variablen X_1 lässt sich wie folgt bestimmen:

$$b_1 = b_{1s} \cdot \frac{s_Y}{s_{X_1}} \quad (\text{F 19.6})$$

Dabei ist

$$b_{1s} = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \quad (\text{F 19.7})$$

das standardisierte Regressionsgewicht, d.h. das Regressionsgewicht, das man erhält, wenn man die Variablen zunächst standardisiert (z -transformiert). Das Regressionsgewicht b_1 lässt sich einfach aus dem standardisierten Regressionsgewicht b_{1s} berechnen, indem

man es mit der Standardabweichung der abhängigen Variablen multipliziert und durch die Standardabweichung der unabhängigen Variablen X_1 dividiert.

Das Regressionsgewicht der zweiten Variablen X_2 erhält man in analoger Weise:

$$b_2 = b_{2s} \cdot \frac{s_Y}{s_{X_2}}$$

und

$$b_{2s} = \frac{r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2}$$

Je nach den Korrelationen der Variablen untereinander kann sich das Regressionsgewicht einer unabhängigen Variablen in einer multiplen Regressionsanalyse deutlich vom Regressionsgewicht dieser Variablen in einer einfachen Regressionsanalyse unterscheiden. Das Regressionsgewicht einer unabhängigen Variablen in der multiplen Regressionsanalyse kann größer oder kleiner als das Regressionsgewicht dieser Variablen in der einfachen Regressionsanalyse, aber auch mit diesem identisch sein. Zudem kann das Regressionsgewicht in der multiplen Regressionsanalyse auch ein anderes Vorzeichen aufweisen als das Regressionsgewicht dieser Variablen in der einfachen Regressionsanalyse. Wir werden dies im Folgenden anhand einiger empirischer Beispiele demonstrieren. Wir werden das Regressionsgewicht einer Variablen in der einfachen Regressionsanalyse b_{einf} und das Regressionsgewicht einer Variablen in der multiplen Regressionsanalyse b_{mult} nennen. Rein formal lassen sich folgende Fälle identifizieren.

$b_{\text{einf}} = b_{\text{mult}}$ Das Regressionsgewicht einer Variablen in der multiplen Regression entspricht dem Regressionsgewicht dieser Variablen in einer einfachen Regression, wenn alle Prädiktorvariablen untereinander unkorreliert sind.

$b_{\text{einf}} > b_{\text{mult}}$ Das Regressionsgewicht der Variablen X_1 in der multiplen Regression ist kleiner als ihr Regressionsgewicht in der einfachen Regression, wenn folgende Beziehung gilt:

$$r_{YX_1} > \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \quad (\text{F 19.8})$$

Diese Beziehung folgt aus Gleichung F 19.7 und der Tatsache, dass das standardisierte Regressionsgewicht

in der einfachen Regressionsanalyse gleich der Produkt-Moment-Korrelation ist (linke Seite von Gleichung F 19.8). Aus Gleichung F 19.8 folgt z. B. für positive Korrelationen $r_{X_1X_2}$, r_{YX_1} und r_{YX_2} :

$$r_{X_1X_2} < \frac{r_{YX_2}}{r_{YX_1}} \quad (\text{F 19.9})$$

Sofern die beiden unabhängigen Variablen miteinander korreliert sind, ist die Gleichung F 19.9 z. B. immer erfüllt, wenn $r_{YX_2} > r_{YX_1}$, da dann der Ausdruck auf der rechten Seite immer größer als 1 ist, die Korrelation auf der linken Seite aber kleiner als 1 sein muss. Wären beide unabhängigen Variablen zu 1 korreliert, wären die Gleichungen F 19.7 und F 19.8 nicht definiert. In diesem Fall wäre es auch sinnlos, beide Variablen in die Regressionsgleichung aufzunehmen, da die zweite Variable keine neue Information enthält. Man sagt dann, dass zwischen beiden Variablen *exakte Kollinearität* vorliegt. Sind beide unabhängigen Variablen unkorreliert, kann Gleichung F 19.8 nie erfüllt sein, da dann in beiden Seiten der Ungleichung die gleiche Korrelation stehen würde, wodurch die Ungleichung falsch wäre.

$b_{\text{einf}} < b_{\text{mult}}$ Das Regressionsgewicht der Variablen X_1 in der multiplen Regression kann auch größer sein als das entsprechende Regressionsgewicht dieser Variablen in der einfachen Regression. Dies ist dann der Fall, wenn folgende Beziehung gilt:

$$r_{YX_1} < \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2} \quad (\text{F 19.10})$$

und somit z. B. für positive Korrelationen $r_{X_1X_2}$, r_{YX_1} und r_{YX_2} :

$$r_{X_1X_2} > \frac{r_{YX_2}}{r_{YX_1}} \quad (\text{F 19.11})$$

Diese Ungleichung ist z. B. immer dann erfüllt, wenn die Variable X_2 mit der Kriteriumsvariablen unkorreliert ist und die beiden unabhängigen Variablen positiv korreliert sind. Diesen speziellen Fall einer sog. klassischen Suppressorvariablen werden wir in Abschnitt 19.8 behandeln. Die Bedingung in Ungleichung F 19.11 zeigt, dass dieser Fall umso eher eintritt, je höher die Korrelation der beiden unabhängigen Variablen ist und je stärker sich die Korrelationen zwischen den unabhängigen und der abhängigen Variablen zugunsten einer höheren Korrelation r_{YX_1} unterscheiden.

Wir haben diese drei verschiedenen Fälle für die Variable X_1 behandelt, sie gelten natürlich in analoger Weise auch für die Variable X_2 . Diese Fälle zeigen, dass der Austausch oder das Weglassen einer einzigen unabhängigen Variablen die Regressionsgewichte der anderen Prädiktoren deutlich verändern kann.

Bestimmung des Achsenabschnitts b_0

Wie bei der einfachen Regression wird der Achsenabschnitt, also die Regressionskonstante b_0 , auf der Grundlage der Mittelwerte und der Regressionsgewichte bestimmt:

$$b_0 = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2 \quad (\text{F 19.12})$$

Die Größe der Regressionskonstanten b_0 hängt von den Mittelwerten der abhängigen und der unabhängigen Variablen sowie von den Regressionsgewichten b_1 und b_2 ab. Die Regressionskonstante b_0 entspricht dem erwarteten Wert der abhängigen Variablen Y , wenn alle unabhängigen Variablen den Wert 0 annehmen. Abbildung 19.3 ist zu entnehmen, dass b_0 der Schnittpunkt der Regressionsebene mit der y -Achse ist. Im Falle standardisierter Variablen ist die Regressionskonstante b_0 immer gleich 0.

19.4 Bedeutung der Regressionsgewichte

Die Bedeutung der Regressionsgewichte kann man auf zwei verschiedene Arten erläutern: Man kann multiple Regressionsgewichte auffassen

- (1) als Regressionsgewichte bedingter einfacher Regressionen oder
- (2) als das Regressionsgewicht zweier Regressionsresiduen.

Wir beschränken uns der Einfachheit halber wieder auf den Fall zweier Prädiktoren.

19.4.1 Multiple Regressionsgewichte als Regressionsgewichte bedingter einfacher Regressionen

Das multiple Regressionsgewicht b_j kann als Regressionsgewicht einer unabhängigen Variablen in einer bedingten einfachen Regressionsanalyse interpretiert

werden. Eine bedingte Regressionsanalyse liegt dann vor, wenn wir die Ausprägungen aller anderen unabhängigen Variablen konstant halten. Dies ist z. B. dann der Fall, wenn nur Personen mit denselben Ausprägungen auf den unabhängigen Variablen betrachtet werden. Im Beispiel der beiden Stimmungsregulationskompetenzen ist das Regressionsgewicht der Fähigkeit zur Aufrechterhaltung guter Stimmung identisch mit dem Regressionsgewicht dieser Variablen in einer einfachen Regressionsanalyse, wenn die Fähigkeit zur Verbesserung schlechter Stimmung auf einen Wert fixiert ist. Das multiple Regressionsgewicht der Fähigkeit zur Aufrechterhaltung guter Stimmung ist also nichts anderes als das einfache Regressionsgewicht dieser Variablen in Subgruppen von Personen, die sich in Bezug auf den anderen Prädiktor nicht unterscheiden. Durch die multiple Regressionsanalyse überprüft man den Einfluss einer unabhängigen Variablen auf die abhängige Variable bei Konstanzhaltung aller anderen unabhängigen Variablen. Dieser Sachverhalt ist für drei ausgewählte Gruppen von Personen, die unterschiedliche Werte auf der Variablen VS aufweisen, in Abbildung 19.4 dargestellt. Man sieht hier, dass die Abhängigkeit des Wohlbefindens von der Fähigkeit zur Aufrechterhaltung guter Stimmung in allen drei Subgruppen gleich ist, dass sich die drei Subgruppen jedoch in ihren Achsenabschnitten unterscheiden. Dies lässt sich dadurch erklären, dass die Fähigkeit zur Verbesserung schlechter Stimmung auch einen Effekt auf das Wohlbefinden hat. Die drei Gleichungen, die in Abbildung 19.4 dargestellt sind, erhält man, indem man einfach die drei Werte (1, 2, 3) für die Verbesserung schlechter Stimmung in die Regressionsgleichung einsetzt:

Regressionsgleichung für VS = 1:

$$\begin{aligned} \widehat{WB} &= 1,801 + 0,354 \cdot AS + 0,336 \cdot 1 \\ &= 2,137 + 0,354 \cdot AS \end{aligned}$$

Regressionsgleichung für VS = 2:

$$\begin{aligned} \widehat{WB} &= 1,801 + 0,354 \cdot AS + 0,336 \cdot 2 \\ &= 2,473 + 0,354 \cdot AS \end{aligned}$$

Regressionsgleichung für VS = 3:

$$\begin{aligned} \widehat{WB} &= 1,801 + 0,354 \cdot AS + 0,336 \cdot 3 \\ &= 2,809 + 0,354 \cdot AS \end{aligned}$$

Diese drei Regressionsgleichungen unterscheiden sich von der Regressionsgleichung, die man erhalten würde,

wenn man die Variable *VS* *nicht* mit berücksichtigen würde (also für den Fall einer einfachen bivariaten Regression von *WB* auf *AS*). Da *AS* und *VS* untereinander korreliert sind ($r = 0,445$), deckt die Fähigkeit zur Aufrechterhaltung guter Stimmung auch einen Teil der Fähigkeit zur Verbesserung schlechter Stimmung ab. Da nun auch die Verbesserung schlechter Stimmung mit dem Wohlbefinden zusammenhängt, ist ein Teil des Effekts von *AS* auf *WB* redundant mit dem Effekt von *VS* auf *WB*. Diese Redundanz wird bei der multiplen Regression berücksichtigt. Das Regressionsgewicht in einer multiplen Regressionsanalyse gibt also nur an, inwieweit eine unabhängige Variable einen Beitrag zur Erklärung bzw. Prädiktion der abhängigen Variablen leisten kann, der über den Erklärungs- bzw. Prädiktionsbeitrag aller anderen unabhängigen Variablen in der Gleichung hinausgeht. Das Regressionsgewicht eines Prädiktors X_j gibt somit an, inwieweit die Variation im Kriterium Y auf diesen Prädiktor zurückzuführen ist unter der Bedingung, dass alle anderen Prädiktorvariablen konstant gehalten werden. Daher kann man multiple Regressionsgewichte auch als bedingte einfache Regressionsgewichte verstehen.

Bezogen auf unser Beispiel: Das Regressionsgewicht der Variablen *AS* zeigt an, um wie viele Einheiten sich die vorhergesagten Werte für *WB* verändern, wenn sich *AS* um eine Einheit erhöht und *VS* konstant bleibt. Nehmen wir z. B. an, eine Person ($m = 1$) hat einen Wert von 1 auf *AS* und einen Wert von 1 auf *VS*. Hieraus ergibt sich aufgrund der Regressionsgleichung ein erwarteter Wohlbefindenswert von $\widehat{wb}_1 = 2,491$. Nun können wir diesen erwarteten Wohlbefindenswert mit dem erwarteten Wohlbefindenswert einer Person ($m = 2$) vergleichen, die sich zwar nicht in *VS*, aber in *AS* von jener Person unterscheidet: Nehmen wir an, die zweite Person habe den Wert 2 auf der Variablen *AS*. Setzt man nun für diese Person den Wert in die Bestimmungsgleichung ein, erhält man einen erwarteten Wohlbefindenswert von $\widehat{wb}_2 = 2,845$. Die Differenz der erwarteten Wohlbefindenswerte der beiden Personen ($\widehat{wb}_2 - \widehat{wb}_1 = 2,845 - 2,491 = 0,354$) entspricht genau dem (unstandardisierten) multiplen Regressionsgewicht von *AS*.

Die Feststellung, dass das Regressionsgewicht einer Variablen angibt, um welchen erwarteten Wert sich die abhängige Variable ändert, wenn man den Wert der unabhängigen Variablen um eine Einheit erhöht, ohne dass man jedoch die Werte auf den anderen unabhän-

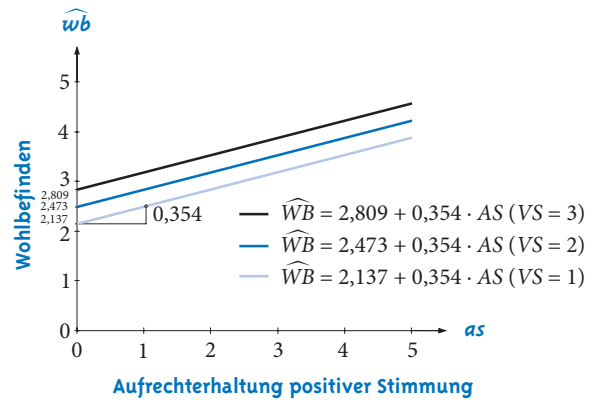


Abbildung 19.4 Bedingte Regression von Wohlbefinden (*WB*) auf Aufrechterhaltung guter Stimmung (*AS*) für drei ausgewählte Werte von Fähigkeit zur Verbesserung schlechter Stimmung (*VS*)

gigen Variablen verändert, gilt für alle unabhängigen Variablen gleichermaßen. Analog zur Abbildung 19.4 könnte man auch die regressive Abhängigkeit des Wohlbefindens von *VS* unter Konstanthaltung von *AS* darstellen. Das Regressionsgewicht der Verbesserung schlechter Stimmung zeigt an, um welchen Wert sich der erwartete \hat{y} -Wert ändert, wenn die Verbesserung schlechter Stimmung um eine Einheit erhöht wird und die Fähigkeit zur Aufrechterhaltung guter Stimmung konstant gehalten wird.

19.4.2 Multiple Regressionsgewichte als Regressionsgewichte von Regressionsresiduen

Die zweite Interpretationsmöglichkeit eines Regressionsgewichts b_j in der multiplen Regressionsanalyse besteht darin, es als Regressionsgewicht zweier Residualvariablen aufzufassen. Die beiden Residualvariablen erhält man, indem man den fraglichen Prädiktor X_j und das Kriterium Y um alle Abhängigkeiten von den anderen Prädiktorvariablen (die in die multiple Regressionsgleichung mit aufgenommen werden) bereinigt. Zur Veranschaulichung greifen wir noch einmal auf das Beispiel zurück, in dem das Wohlbefinden auf Unterschiede in beiden Stimmungsregulationskompetenzen zurückgeführt wird. Wir wollen die Bedeutung des multiplen Regressionsgewichts für die Variable X_1 erläutern. Sie trifft in analoger Weise auf die Variable X_2

zu. Die Bedeutung des Regressionsgewichts ist eng mit der Partialkorrelation verknüpft, weswegen die multiplen Regressionsgewichte auch *Partialregressionsgewichte* genannt werden.

In Abbildung 19.5 sind zwei einfache Regressionsanalysen dargestellt, in denen sowohl Unterschiede im Wohlbefinden (Y) als auch Unterschiede in der Aufrechterhaltung guter Stimmung (X_1) auf Unterschiede in der Verbesserung schlechter Stimmung (X_2) zurückgeführt werden. Da die Verbesserung schlechter Stimmung die Unterschiede im Wohlbefinden und in der Aufrechterhaltung guter Stimmung nicht perfekt erklären kann, gibt es zwei Residualvariablen: die Residualvariable $E_{Y(X_2)}$, deren Werte die Abweichungen der beobachteten y -Werte (Wohlbefinden) von den aufgrund der Variablen X_2 (Verbesserung schlechter Stimmung) vorhergesagten \hat{y} -Werten sind, und die Residualvariable $E_{X_1(X_2)}$, deren Werte die Abweichungen der beobachteten x_1 -Werte (Aufrechterhaltung guter Stimmung) von den aufgrund der Variablen X_2 (Verbesserung schlechter Stimmung) vorhergesagten \hat{x}_1 -Werten sind. Beide Residualvariablen sind also nun um den Einfluss von X_2 bereinigt; anders gesagt: Der Einfluss von X_2 wurde sowohl aus X_1 als auch aus Y auspartialisiert. Trotzdem können die beiden Residualvariablen miteinander korrelieren: Die Fähigkeit zur Aufrechterhaltung guter Stimmung kann mit mehr Wohlbefinden einhergehen, obwohl man aus diesem Zusammenhang die Fähigkeit zur Verbesserung schlechter Stimmung auspartialisiert hat. Eine solche mögliche Kovariation zwischen $E_{Y(X_2)}$ und $E_{X_1(X_2)}$ ist in Abbildung 19.5 a durch den Kovarianzbogen dargestellt.

Berechnet man nun eine einfache lineare Regressionsanalyse mit $E_{Y(X_2)}$ als Kriterium und $E_{X_1(X_2)}$ als Prädiktor, so erhält man ein einfaches Regressionsgewicht, das in der Abbildung 19.5 b als b_1 bezeichnet wird. Dieses Regressionsgewicht b_1 der Regressionsresiduen entspricht genau dem multiplen Regressionsgewicht der Variablen X_1 in der multiplen Regressionsanalyse. Das Regressionsgewicht b_1 kennzeichnet also den Teil des Einflusses der unabhängigen Variablen X_1 auf die abhängige Variable Y , der nicht bereits durch X_2 erklärt wird. Das Regressionsgewicht b_1 ist also ein Maß dafür, welchen Einfluss die Variable X_1 über die Variable X_2 hinausgehend auf die Variable Y hat. In analoger Weise lässt sich das Partialregressionsgewicht für die Variable X_2 bestimmen.

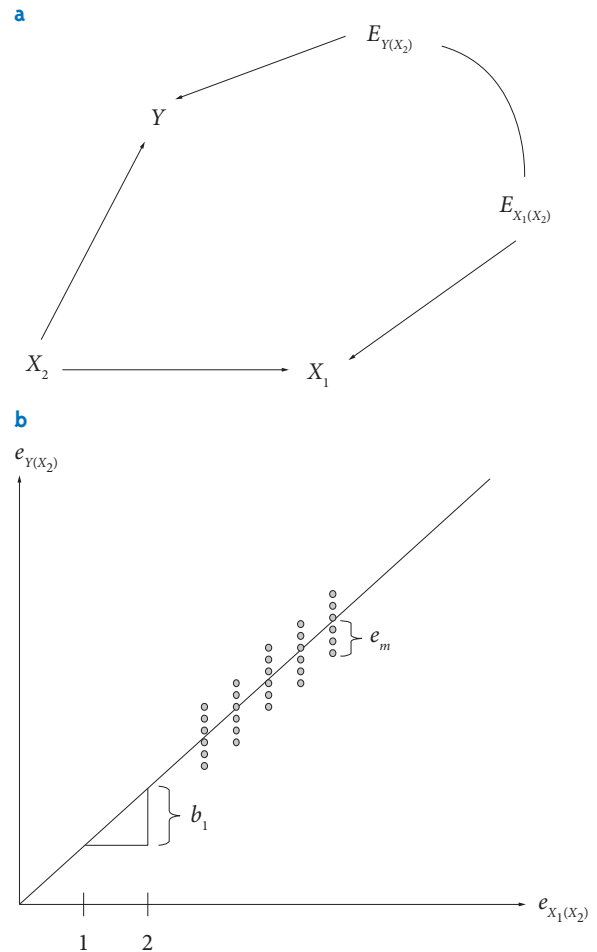


Abbildung 19.5 Partialregressionsgewicht b_1 als Regressionsgewicht von Regressionsresiduen

Vergleich mit der Partialkorrelation

Die Nähe zur Partialkorrelation wird hier deutlich, da die Partialkorrelation als die bivariate Korrelation von Regressionsresiduen definiert ist (s. Abschn. 18.2). Es gibt jedoch einen bedeutsamen Unterschied zwischen dem Partialregressionsgewicht und der Partialkorrelation. Das Partialregressionsgewicht ist das unstandardisierte Regressionsgewicht zweier Residualvariablen. Transformiert man die X - und Y -Variablen so, dass sie einen Mittelwert von 0 und eine Standardabweichung von 1 haben (z -Transformation), erhält man ein standardisiertes multiples Regressionsgewicht, dessen Berechnung bereits dargestellt wurde (s. Gleichung F 19.7). Dieses standardisierte Regressionsgewicht entspricht jedoch nicht der Partialkorrelation, da im Falle standardisierter

Variablen die X- und Y-Variablen standardisiert werden, im Falle der Partialkorrelation hingegen die Residualvariablen. Partialregressionsgewicht und Partialkorrelation sind verwandt, aber nicht identisch. Zwischen den Partialregressionsgewichten und den Partialkorrelationen besteht folgende Beziehung (s. Übung 1):

$$b_1 = r_{X_1 Y \cdot X_2} \cdot \frac{\sqrt{s_Y^2 \cdot (1 - r_{X_2 Y}^2)}}{\sqrt{s_{X_1}^2 \cdot (1 - r_{X_2 X_1}^2)}}$$

bzw.

$$b_2 = r_{X_2 Y \cdot X_1} \cdot \frac{\sqrt{s_Y^2 \cdot (1 - r_{X_1 Y}^2)}}{\sqrt{s_{X_2}^2 \cdot (1 - r_{X_1 X_2}^2)}} \quad (\text{F 19.13})$$

Der Ausdruck $s_Y^2 \cdot (1 - r_{X_2 Y}^2)$ ist die Varianz der Residualvariablen $E_{Y(X_2)}$, $s_{X_1}^2 \cdot (1 - r_{X_2 X_1}^2)$ die Varianz der Residualvariablen $E_{X_1(X_2)}$ und $s_{X_2}^2 \cdot (1 - r_{X_1 X_2}^2)$ die Varianz der Residualvariablen $E_{X_2(X_1)}$. Gleichung F 19.13 zeigt die strukturelle Ähnlichkeit zum Regressionskoeffizienten in der einfachen Regressionsanalyse (s. Gleichung F 17.4). Das Regressionsgewicht in der einfachen Regressionsanalyse wird bestimmt als die Korrelation nullter Ordnung multipliziert mit dem Verhältnis aus den Standardabweichungen der abhängigen Variablen und der unabhängigen Variablen. Bei zwei Prädiktoren ist der Partialregressionskoeffizient eines Prädiktors das Produkt aus der Partialkorrelation erster Ordnung dieses Prädiktors mit dem Kriterium, multipliziert mit dem Verhältnis der Standardabweichungen des Kriteriums und dieses Prädiktors, nachdem aus beiden Variablen der Einfluss des anderen Prädiktors auspartialisiert wurde.

19.4.3 Unstandardisierte vs. standardisierte Regressionsgewichte

Wir haben bereits bei der Bestimmung der Regressionsgewichte in Abschnitt 17.7 gesehen, dass zwei Arten von Regressionsgewichten unterschieden werden können: Unstandardisierte Regressionsgewichte werden auf der Grundlage der Metrik der Variablen berechnet, in der sie gemessen wurden. Darüber hinaus lassen sich standardisierte Regressionsgewichte als Regressionsgewichte von z-transformierten Variablen

bestimmen. Das unstandardisierte Regressionsgewicht gibt an, um welchen Wert sich die vorhergesagten \hat{y} -Werte ändern, wenn die unabhängige Variable sich um eine Einheit ändert. Das standardisierte Regressionsgewicht gibt in analoger Weise an, um wie viele Standardabweichungseinheiten sich die vorhergesagten \hat{y} -Werte ändern, wenn sich die unabhängige Variable um eine Standardabweichungseinheit verändert. Üblicherweise gibt man bei einer Regressionsanalyse beide Regressionsgewichte an. Standardisierte und unstandardisierte Regressionsgewichte sind für unterschiedliche Fragestellungen geeignet, wie wir bereits bei der einfachen Regressionsanalyse in Abschnitt 17.7 besprochen haben und hier nochmals zusammenfassend wiederholen wollen.

Vergleich verschiedener Gruppen

Geht es darum, verschiedene Gruppen von Personen in Bezug auf die regressive Abhängigkeit zu vergleichen, so greift man auf die unstandardisierten Regressionsgewichte zurück – sofern in den verschiedenen Gruppen dieselben Messinstrumente zum Einsatz kamen. Ist man z.B. daran interessiert, ob sich Frauen und Männer in der Beziehung zwischen Wohlbefinden und den beiden Stimmungsregulationskompetenzen unterscheiden, so würde man statistisch die Gleichheit der unstandardisierten Regressionsgewichte zwischen beiden Gruppen überprüfen. Für diese Fragestellung sind die unstandardisierten Regressionsgewichte den standardisierten Regressionsgewichten vorzuziehen, da die Frage im Mittelpunkt steht, ob Frauen und Männer, die über dieselben Regulationskompetenzen verfügen, auch denselben erwarteten Wohlbefindenswert haben. Dies wäre dann der Fall, wenn die Regressionskoeffizienten in beiden Geschlechtsgruppen gleich wären. Die standardisierten Regressionsgewichte sind hierzu nicht geeignet, da sich Frauen und Männer in den Varianzen der Variablen unterscheiden können. Ist dies der Fall, so werden in beiden Geschlechtsgruppen unterschiedliche Standardisierungen vorgenommen, da die Abweichungswerte durch unterschiedliche Standardabweichungen geteilt werden. Dies hat zur Folge, dass sich trotz gleicher unstandardisierter Regressionsgewichte unterschiedliche standardisierte Regressionsgewichte ergeben können. Für die Prognose des Verhaltens in beiden Geschlechtsgruppen sind jedoch die Variablen in ihrer ursprünglichen Metrik ausschlaggebend.

Vergleich verschiedener Variablen

Auf standardisierte Regressionsgewichte greift man zurück, wenn man verschiedene Variablen, die in einer unterschiedlichen Metrik erfasst wurden, hinsichtlich ihres Vorhersagebeitrages an der abhängigen Variablen miteinander vergleichen möchte. Für diesen Vergleich sind die unstandardisierten Regressionsgewichte weniger gut geeignet, da die Größe eines unstandardisierten Regressionsgewichts auch von der Metrik einer Variablen abhängt. Dies haben wir schon bei der einfachen Regressionsanalyse gesehen. Hat man z. B. in verschiedenen Studien die regressive Abhängigkeit des Wohlbefindens von der Stimmungsregulationskompetenz untersucht und die Merkmale mit unterschiedlichen Messinstrumenten erfasst, sind die Regressionsgewichte aufgrund der Unterschiede in der Metrik zwischen den Studien nur schwer zu vergleichen. Standardisiert man die Variablen in den verschiedenen Studien durch eine z-Transformation, lassen sich die Regressionsgewichte zwischen den Studien vergleichen. Das standardisierte Regressionsgewicht zeigt die erwartete Veränderung der \hat{y} -Werte an, wenn sich die unabhängige Variable um eine Standardabweichungseinheit verändert und die anderen unabhängigen Variablen konstant gehalten werden.

19.5 Lineare Regression für mehrere metrische unabhängige Variablen

Im Falle von mehr als zwei unabhängigen Variablen erweitert sich die Regressionsgleichung durch die Hinzunahme der mit einem (Partial-)Regressionsgewicht gewichteten Werte der weiteren Variablen X_j :

$$\hat{y}_m = b_0 + b_1 \cdot x_{m1} + b_2 \cdot x_{m2} + \dots + b_j \cdot x_{mj} + \dots + b_k \cdot x_{mk}$$

Die Bedeutung der Regressionskoeffizienten lässt sich in Analogie zum Fall zweier unabhängiger Variablen herleiten. Der Achsenabschnitt b_0 (die additive Konstante) ergibt sich aus den Mittelwerten der abhängigen und der mit den Regressionsgewichten gewichteten unabhängigen Variablen:

$$b_0 = \bar{y} - (b_1 \cdot \bar{x}_1 + b_2 \cdot \bar{x}_2 + \dots + b_j \cdot \bar{x}_j + \dots + b_k \cdot \bar{x}_k) \quad (\text{F 19.14})$$

Ein Regressionsgewicht b_j erhält man in Analogie zum Falle zweier Variablen wie folgt:

- ▶ Aus der Variablen X_j wird der Einfluss aller anderen unabhängigen Variablen $X_{j'} (j' \neq j)$ auspartialisiert. Man erhält dann eine Residualvariable $E_{X_j(X_{\setminus j})}$, die den Teil von X_j repräsentiert, der nicht durch alle anderen unabhängigen Variablen gemeinsam vorhergesagt werden kann. Die Schreibweise $X_{\setminus j}$ nutzen wir, um anzuzeigen, dass alle k unabhängigen Variablen mit Ausnahme der Variable X_j betrachtet werden.
- ▶ Aus der abhängigen Variablen Y wird ebenfalls der Einfluss aller anderen unabhängigen Variablen $X_{j'} (j' \neq j)$ auspartialisiert, sodass man eine Residualvariable $E_Y(X_{\setminus j})$ erhält, die den Teil von Y repräsentiert, der nicht durch die unabhängigen Variablen (X_j ausgenommen) gemeinsam vorhergesagt werden kann.
- ▶ Das Regressionsgewicht b_j ist das Regressionsgewicht einer einfachen linearen Regressionsanalyse, in der $E_Y(X_{\setminus j})$ auf $E_{X_j(X_{\setminus j})}$ zurückgeführt wird.

Das Regressionsgewicht b_j hängt in Analogie zum Fall zweier Prädiktoren wie folgt mit der Partialkorrelation zusammen:

$$b_j = r_{X_j Y \cdot X_{\setminus j}} \cdot \frac{\sqrt{s_Y^2 \cdot (1 - R_{Y|X_{\setminus j}}^2)}}{\sqrt{s_{X_j}^2 \cdot (1 - R_{X_j|X_{\setminus j}}^2)}} \quad (\text{F 19.15})$$

In dieser Formel bezeichnet $r_{X_j Y \cdot X_{\setminus j}}$ die Partialkorrelation ($k-1$ -ter Ordnung zwischen Y und X_j , d.h. die Korrelation der Regressionsresiduen, nachdem sowohl aus Y als auch aus X_j der Teil auspartialisiert wurde, der auf die anderen unabhängigen Variablen zurückgeführt werden kann. Die Abhängigkeit des Partialregressionskoeffizienten b_j in der multiplen Regression von der Partialkorrelation ($k-1$ -ter Ordnung zeigt, was es bedeutet, dass in der multiplen Regression der Einfluss anderer Variablen kontrolliert wird. Es wird überprüft, ob die beiden Variablen Y und X_j auch dann noch kovariieren, wenn der Einfluss aller anderen Variablen durch Auspartialisieren statistisch kontrolliert wurde. Das Regressionsgewicht b_j ist immer dann 0, wenn die Partialkorrelation gleich 0 ist. Auch das Prinzip der Berücksichtigung von Redundanzen wird deutlich. Wenn X_j keine Information für die Prädiktion von Y enthält, die nicht schon durch die anderen unabhängigen Variablen abgedeckt wird, dann ist das Partialregressionsgewicht b_j gleich 0.

Der Ausdruck $s_Y^2 \cdot (1 - R_{Y|X_{\setminus j}}^2)$ bezeichnet die Residualvarianz von Y , d.h. die Varianz von Y , die übrig bleibt, wenn die durch alle unabhängigen Variablen X (mit Ausnahme von X_j) determinierte Varianz entfernt wurde. Mit $R_{Y|X_{\setminus j}}^2$ wird der Determinationskoeffizient bezeichnet, d.h. der Anteil der Varianz von Y , der durch alle Prädiktoren mit Ausnahme von X_j determiniert wird. Es ist die quadrierte multiple Korrelation zwischen der abhängigen Variablen Y und allen unabhängigen Variablen ausgenommen X_j . Was sich dahinter verbirgt, werden wir im nächsten Abschnitt erläutern. Das Partialregressionsgewicht b_j lässt sich nicht mehr so einfach berechnen wie im Falle zweier unabhängiger Variablen. Zu seiner Bestimmung greift man auf die matrixalgebraische Bestimmungsformel zurück (s. Anhang B).

19.6 Multiple Korrelation und Determinationskoeffizient

Wie bei der einfachen Regressionsanalyse lässt sich auch bei der multiplen Regressionsanalyse die Stärke des Zusammenhangs zwischen der abhängigen Variablen und den unabhängigen Variablen quantifizieren. Die Pendant zum einfachen Korrelationskoeffizienten und zum Determinationskoeffizienten der einfachen Regressionsanalyse sind der multiple Korrelationskoeffizient und der multiple Determinationskoeffizient R^2 . Im Gegensatz zur Produkt-Moment-Korrelation, die mit r symbolisiert wird, bezeichnet man die multiple Korrelation mit einem großen R . Die multiple Korrelation ist die Korrelation zwischen den beobachteten y -Werten und den durch die Regressionsanalyse vorhergesagten \hat{y} -Werten:

$$R = r_{Y\hat{Y}} \quad (\text{F 19.16})$$

Da die Variable \hat{Y} eine Linearkombination aller unabhängigen Variablen X_1, \dots, X_k ist, lässt sich die multiple Korrelation auch wie folgt ausdrücken:

$$R = r_Y(b_0 + b_1X_1 + b_2X_2 + \dots + b_jX_j + \dots + b_kX_k) \quad (\text{F 19.17})$$

Wir symbolisieren die multiple Korrelation im Folgenden mit $R_{Y|X_1, X_2, \dots, X_j, \dots, X_k}$, um kenntlich zu machen, mit welcher Menge von unabhängigen Variablen X_j die abhängige Variable Y korreliert ist. Die multiple Korrelation kann nicht negativ werden, da die Regressionsgewichte so bestimmt werden, dass die Abweichung der

vorhergesagten Kriteriumswerte von den beobachteten Kriteriumswerten minimal wird. Folglich gehen höhere y -Werte tendenziell immer mit höheren \hat{y} -Werten einher und geringere y -Werte tendenziell immer mit geringeren \hat{y} -Werten. Somit kann die multiple Korrelation nur zwischen 0 und 1 variieren.

Das Quadrat der multiplen Korrelation ist der multiple Determinationskoeffizient R^2 :

$$R^2 = r_{Y\hat{Y}}^2 \quad (\text{F 19.18})$$

Hieraus folgt, dass der multiple Determinationskoeffizient gleich dem Verhältnis aus der Varianz der vorhergesagten \hat{y} -Werte und der beobachteten y -Werte ist (s. Übung 2):

$$R^2 = \frac{s_{\hat{Y}}^2}{s_Y^2} \quad (\text{F 19.19})$$

Der multiple Determinationskoeffizient (das multiple Bestimmtheitsmaß) ist der Anteil an der Varianz der abhängigen Variablen, der durch alle unabhängigen Variablen gemeinsam bestimmt (determiniert) wird. Der Determinationskoeffizient kann Werte zwischen 0 und 1 annehmen. Ein Determinationskoeffizient von 0 bedeutet, dass die Varianz der vorhergesagten Variablen \hat{Y} gleich 0 ist; d.h., unabhängig von ihren Werten auf den unabhängigen Variablen erhält jede Person denselben vorhergesagten Wert. Die unabhängigen Variablen sind in diesem Fall ungeeignet, Unterschiede in der abhängigen Variablen vorherzusagen. Ein Determinationskoeffizient von 1 besagt, dass alle vorhergesagten Werte mit allen beobachteten Werten übereinstimmen. Die unabhängigen Variablen erklären in diesem Fall die gesamte Variation der abhängigen Variablen. Multiple Determinationskoeffizienten zwischen 0 und 1 zeigen das Maß der Vorhersageleistung der unabhängigen Variablen an. Je höher der Determinationskoeffizient ist, umso größer ist der Anteil der durch die unabhängigen Variablen erklärten Varianz der abhängigen Variablen.

Varianzzerlegung

Die Bedeutung des Determinationskoeffizienten kann man sich auch an der Zerlegung der Varianz der beobachteten y -Werte in die Varianz der vorhergesagten Werte und die Varianz der Residualwerte klarmachen:

$$s_Y^2 = s_{\hat{Y}}^2 + s_E^2 \quad (\text{F 19.20})$$

Teilt man beide Seiten der Gleichung F 19.20 durch die Varianz der abhängigen Variablen, so erhält man:

$$\frac{s_Y^2}{s_Y^2} = 1 = \frac{s_Y^2}{s_Y^2} + \frac{s_E^2}{s_Y^2} \quad (\text{F 19.21})$$

Hieraus folgt, dass sich die beiden Varianzverhältnisse auf der rechten Seite der Gleichung F 19.21 zu 1 addieren. Der erste dieser beiden Varianzquotienten ist der multiple Determinationskoeffizient. Zieht man den multiplen Determinationskoeffizienten von 1 ab, erhält man den multiplen Indeterminationskoeffizienten (s. Abschn. 17.5):

$$1 - R^2 = \frac{s_E^2}{s_Y^2}$$

Quadratsummenzerlegung

Wie wir bereits in Abschnitt 17.4 gesehen haben, korrespondiert zur Varianzzerlegung in Formel F 19.20 die Quadratsummenzerlegung:

$$\sum_{m=1}^n (y_m - \bar{y})^2 = \sum_{m=1}^n (y_m - \hat{y}_m)^2 + \sum_{m=1}^n (\hat{y}_m - \bar{y})^2 \quad (\text{F 19.22})$$

Damit lässt sich der multiple Determinationskoeffizient wie folgt definieren:

$$R^2 = \frac{s_Y^2}{s_Y^2} = \frac{\sum_{m=1}^n (\hat{y}_m - \bar{y})^2}{\sum_{m=1}^n (y_m - \bar{y})^2} \quad (\text{F 19.23})$$

Der multiple Indeterminationskoeffizient lässt sich dann wie folgt definieren:

$$1 - R^2 = \frac{s_E^2}{s_Y^2} = \frac{\sum_{m=1}^n (y_m - \hat{y}_m)^2}{\sum_{m=1}^n (y_m - \bar{y})^2} \quad (\text{F 19.24})$$

Der multiple Indeterminationskoeffizient gibt den Anteil der unerklärten Varianz in Y an. Da sich der multiple Determinationskoeffizient und der Indeterminationskoeffizient zu 1 addieren, reicht es aus, einen von beiden zu berichten. Üblicherweise wird der multiple Determinationskoeffizient berichtet.

Multiple Determination als Summe von Semipartialdeterminationen zunehmend höherer Ordnung

Der multiple Determinationskoeffizient lässt sich auch als Summe von Semipartialdeterminationen (quadratischen Semipartialkorrelationen) zunehmend höherer Ordnung darstellen. Dies wollen wir wieder anhand eines empirischen Beispiels illustrieren. Wir greifen hierbei auf die abhängige Variable Wohlbefinden (Y) und die drei unabhängigen Variablen Fähigkeit zur Aufrechterhaltung guter Stimmung (X_1), Fähigkeit zur Verbesserung schlechter Stimmung (X_2) und Extraversion (X_3) zurück. Wir betrachten nun drei Regressionsanalysen:

- (1) die einfache Regressionsanalyse mit der abhängigen Variablen Wohlbefinden (Y) und der unabhängigen Variablen Aufrechterhaltung guter Stimmung (X_1) – Modell I
- (2) die multiple Regressionsanalyse mit der abhängigen Variablen Wohlbefinden (Y) und den unabhängigen Variablen Aufrechterhaltung guter Stimmung (X_1) und Verbesserung schlechter Stimmung (X_2) – Modell II
- (3) die multiple Regressionsanalyse mit der abhängigen Variablen Wohlbefinden (Y) und den unabhängigen Variablen Aufrechterhaltung guter Stimmung (X_1), Verbesserung schlechter Stimmung (X_2) und Extraversion (X_3) – Modell III

Die Determinationskoeffizienten für diese drei Regressionsanalysen sind in Tabelle 19.1 angegeben. Die Ergebnisse im Folgenden beziehen sich auf die Daten von denjenigen $n = 237$ Personen, von denen Angaben auf allen vier Variablen vorliegen. Diese Tabelle enthält auch die Änderungen in den Determinationskoeffizienten, wenn man von Modell I zu Modell II und von Modell II zu Modell III übergeht. Wir bezeichnen solche Differenzen in R^2 mit dem griechischen Großbuchstaben Δ (Delta). Die durch einen hinzugenommenen Prädiktor zusätzlich zu den bisherigen Prädiktoren aufgeklärte Varianz in Y bezeichnet man auch als *Inkrement* in R^2 . Zusätzliche Prädiktoren klären also inkrementelle Varianz in Y auf.

Der Determinationskoeffizient in Modell I entspricht der quadrierten Produkt-Moment-Korrelation zwischen dem Wohlbefinden und der Aufrechterhaltung guter Stimmung ($R_I^2 = 0,249$). Der Determinationskoeffizient in Modell II ($R_{II}^2 = 0,338$) entspricht der

Tabelle 19.1 Determinationskoeffizienten und Veränderungen der Determinationskoeffizienten in drei Regressionsmodellen

Modell	Unabhängige Variable(n)	R^2	ΔR^2
I	Aufrechterhaltung guter Stimmung	0,249	
II	Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung	0,338	0,089
III	Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung, Extraversion	0,339	0,001

quadrierten multiplen Korrelation zwischen dem Wohlbefinden und den beiden Stimmungsregulationskompetenzen. Die Veränderung in R^2 zwischen Modell I und Modell II ($\Delta R_{II}^2 = 0,089$) entspricht dem quadrierten Semipartialkorrelationskoeffizienten zwischen dem Wohlbefinden und der Verbesserung schlechter Stimmung, nachdem aus der Verbesserung schlechter Stimmung die Aufrechterhaltung guter Stimmung auspartialisiert wurde. Der multiple Determinationskoeffizient in Modell III ($R_{III}^2 = 0,339$) entspricht der quadrierten multiplen Korrelation zwischen dem Wohlbefinden und den drei unabhängigen Variablen. Die Veränderung in R^2 zwischen Modell II und Modell III ($\Delta R_{III}^2 = 0,001$) entspricht dem quadrierten Semipartialkorrelationskoeffizienten zwischen dem Wohlbefinden und der Extraversion, nachdem aus der Extraversion der Einfluss der beiden Stimmungsregulationskompetenzen auspartialisiert wurde. Die multiple Determination lässt sich somit als Summe von Semipartialdeterminationen zunehmend höherer Ordnung darstellen:

$$R_{Y|X_1, X_2, X_3}^2 = r_{X_1 Y}^2 + r_{(X_2 \cdot X_1) Y}^2 + r_{(X_3 \cdot X_1, X_2) Y}^2 \quad (\text{F 19.25})$$

Das schrittweise Vorgehen kann man sich somit so vorstellen:

- (1) Zunächst wird der erste Prädiktor (X_1) mit dem Kriterium korreliert. Das Quadrat der Korrelation entspricht dem Varianzanteil des Kriteriums Y , den dieser erste Prädiktor (X_1) erklären kann.
- (2) Dann wird der erste Prädiktor (X_1) aus dem zweiten Prädiktor (X_2) auspartialisiert. Die Residualvariable

$E_{X_2(X_1)}$ wird wiederum mit dem Kriterium Y korreliert. Das Quadrat der Semipartialkorrelation erster Ordnung ($r_{(X_2 \cdot X_1) Y}^2$) gibt den Varianzanteil des Kriteriums an, den der zweite Prädiktor zusätzlich zum ersten Prädiktor erklärt.

- (3) Dann werden die beiden ersten Prädiktoren aus einem dritten auspartialisiert und die Residualvariable $E_{X_3(X_1, X_2)}$ mit dem Kriterium Y korreliert. Das Quadrat der Semipartialkorrelation zweiter Ordnung ($r_{(X_3 \cdot X_1, X_2) Y}^2$) gibt den Varianzanteil des Kriteriums wieder, den der dritte Prädiktor zusätzlich zu den beiden ersten Prädiktoren erklärt.
- (4) Die Summe aller so berechneten Semipartialdeterminationen entspricht der multiplen Determination (dem Determinationskoeffizienten in der multiplen Regressionsanalyse).

Die Reihenfolge, in der die unabhängigen Variablen in die Regressionsgleichung aufgenommen werden, spielt für die Gesamtdetermination keine Rolle. Die Summe der Semipartialdeterminationen ist für eine Menge von unabhängigen Variablen immer gleich. Allerdings hängt die Höhe der Semipartialdetermination einer unabhängigen Variablen davon ab, welche Prädiktoren zuvor berücksichtigt wurden. Dies lässt sich wiederum anhand unseres Beispiels einfach veranschaulichen. In Tabelle 19.2 wurde die Reihenfolge der Prädiktoren verändert. In Modell I ist nur die Extraversion als unabhängige Variable erhalten. Der Determinationskoeffizient beträgt $R_I^2 = 0,063$, d. h., die Extraversion erklärt 6,3 % der Varianz des Wohlbefindens. Im Modell II wurde zusätzlich die Aufrechterhaltung guter Stimmung aufgenommen. Der Determinationskoeffizient beträgt $R_{II}^2 = 0,254$. Dies bedeutet, dass die Aufrechterhaltung guter Stimmung zusätzlich zur Extraversion 19,1 % der Varianz des Wohlbefindens erklärt. Nach Auspartialisierung der Extraversion beträgt der quadrierte Semipartialkorrelationskoeffizient zwischen dem Wohlbefinden und der Aufrechterhaltung guter Stimmung somit 0,191. Im Modell III wurde als dritte Prädiktorvariable die Verbesserung schlechter Stimmung mit in die Regressionsgleichung aufgenommen. Der Determinationskoeffizient beträgt $R_{III}^2 = 0,339$ und ist identisch mit R_{III}^2 in Tabelle 19.1. Durch die Verbesserung schlechter Stimmung wird über die Extraversion und die Aufrechterhaltung guter Stimmung hinaus 8,5 % der Varianz des Wohlbefindens aufgeklärt. Folglich beträgt der quadrierte Semipartialkorrelationsko-

Tabelle 19.2 Determinationskoeffizienten und Veränderungen der Determinationskoeffizienten in drei Regressionsmodellen bei veränderter Reihenfolge der unabhängigen Variablen

Modell	Unabhängige Variable(n)	R^2	ΔR^2
I	Extraversion	0,063	
II	Extraversion, Aufrechterhaltung guter Stimmung	0,254	0,191
III	Extraversion, Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung	0,339	0,085

effizient zweiter Ordnung zwischen der Verbesserung schlechter Stimmung und dem Wohlbefinden 0,085. Der Erklärungsbeitrag einer Variablen hängt somit davon ab, an welcher Stelle sie in die Regressionsgleichung aufgenommen wird. So erklärt z. B. die Aufrechterhaltung guter Stimmung 24,9 % der Varianz des Kriteriums, wenn außer ihr keine andere Prädiktorvariable berücksichtigt wird (Modell I in Tab. 19.1). Hingegen erklärt die Aufrechterhaltung guter Stimmung nur 19,1 % der Varianz des Kriteriums, wenn gleichzeitig auch die Extraversion als Prädiktor berücksichtigt wird (Modell II in Tab. 19.2).

Nützlichkeit eines Prädiktors

Die Semipartialdetermination der höchstmöglichen Ordnung hat eine ganz spezielle Bedeutung. Sie gibt an, wie viel Varianz ein Prädiktor zusätzlich zu *allen anderen* Prädiktoren erklärt, und beziffert somit die Nützlichkeit des Prädiktors. So ist die Nützlichkeit der Extraversion 0,001 (s. Tab. 19.1), während die Nützlichkeit der Verbesserung schlechter Stimmung 0,085 beträgt (s. Tab. 19.2). Die Nützlichkeit der Aufrechterhaltung guter Stimmung beträgt 0,084 (s. Tab. 19.3). Sie ist nur aufgrund des Rundungsfehlers nicht gleich der Differenz aus 0,339 und 0,256. Die Nützlichkeit der Extraversion ist somit im Vergleich zu den beiden Stimmungsregulationskompetenzen sehr gering. Deren Nützlichkeit ist ungefähr gleich groß (Varianzanteil von 8,4 bzw. 8,5 %).

Semipartialkorrelationskoeffizient vs. Partialregressionskoeffizient

Man beachte den Unterschied zwischen dem Semipartialkorrelationskoeffizienten und dem Partialregres-

Tabelle 19.3 Determinationskoeffizienten und Veränderungen der Determinationskoeffizienten in drei Regressionsmodellen bei veränderter Reihenfolge der unabhängigen Variablen

Modell	Unabhängige Variable(n)	R^2	ΔR^2
I	Verbesserung schlechter Stimmung	0,239	
II	Verbesserung schlechter Stimmung, Extraversion,	0,256	0,017
III	Verbesserung schlechter Stimmung, Extraversion, Aufrechterhaltung guter Stimmung	0,339	0,084

sionskoeffizienten: Wenn man den Regressionskoeffizienten eines Prädiktors bestimmt, partialisiert man alle anderen Prädiktoren sowohl aus dem Kriterium als auch aus dem fraglichen Prädiktor aus und berechnet die einfache Regression von der Residualvariablen des Kriteriums auf die Residualvariable des fraglichen Prädiktors. Bei der Semipartialkorrelation partialisiert man Prädiktoren nur aus den anderen Prädiktoren heraus, nicht aber aus dem Kriterium.

19.7 Inferenzstatistik zur multiplen Regressionsanalyse

Bei der multiplen Regressionsanalyse handelt es sich um eine Erweiterung der einfachen Regressionsanalyse, die wir in Kapitel 17 kennengelernt haben. Daher können die dort behandelten Aspekte auch auf die multiple Regressionsanalyse übertragen werden. Insbesondere lassen sich die beiden in Abschnitt 17.9.2 behandelten Modelle, und zwar das Modell mit deterministischem und das Modell mit stochastischem Regressor, auch für die multiple Regression formulieren. Auch lässt sich das Populationsmodell, das wir in Abschnitt 17.9.1 behandelt haben, auf den Fall mehrerer unabhängiger Variablen erweitern.

19.7.1 Populationsmodell der multiplen Regression

Im Populationsmodell der multiplen Regressionsanalyse geht man davon aus, dass sich die bedingten Er-

wartungswerte der Variablen Y vollständig durch die unabhängigen Variablen und ihre Linearkombination vorhersagen lassen:

$$E(Y | X_1, \dots, X_k) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_j \cdot X_j + \dots + \beta_k \cdot X_k \quad (\text{F 19.26})$$

Der Ausdruck $E(Y | X_1, \dots, X_k)$ bezeichnet die bedingte Erwartung der Variablen Y gegeben alle unabhängigen Variablen X_1, \dots, X_k . Hieraus ergibt sich die Zerlegung der abhängigen Variablen Y :

$$Y = E(Y | X_1, \dots, X_k) + \varepsilon = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_j \cdot X_j + \dots + \beta_k \cdot X_k + \varepsilon \quad (\text{F 19.27})$$

19.7.2 Inferenzstatistische Schätzung und Testung

Um die Parameter des Regressionsmodells inferenzstatistisch absichern und die Konfidenzintervalle schätzen zu können, müssen zusätzliche Annahmen getroffen werden, die wir in Kapitel 17 ausführlich behandelt haben. Wir haben diese Annahmen für das Modell mit einem deterministischen und das Modell mit einem stochastischen Regressor getrennt behandelt. Bei der multiplen Regressionsanalyse werden dieselben Annahmen wie bei der einfachen Regressionsanalyse getroffen. Auch bei der multiplen Regressionsanalyse unterscheiden sich die Parameterschätzungen nicht zwischen beiden Modellen, wohl aber die Stichprobenkennwerteverteilungen des Determinationskoeffizienten, falls dieser in der Population von 0 verschieden ist. Hieraus ergeben sich auch Unterschiede für die Konfidenzintervalle und die Bestimmung der optimalen Stichprobengröße. Wir werden die in Abschnitt 17.9 ausführlich behandelten Annahmen daher nur zusammenfassend wiederholen und auf das Modell der multiplen Regressionsanalyse übertragen.

Modell mit deterministischen Regressoren

Wie wir bereits gesehen haben, wird bei diesem Modell angenommen, dass die Ausprägungen der unabhängigen Variablen X feste Werte sind, also vollständig der Kontrolle des Forschenden unterliegen. Man geht somit von deterministischen unabhängigen Variablen aus, wie es für geplante Untersuchungen zutrifft, in denen man für

festen Werte der unabhängigen Variablen Stichproben auswählt, die sich dann bezüglich ihrer y -Werte unterscheiden können. Dieses Modell wird auch Allgemeines Lineares Modell (ALM) genannt. Wir werden in den Abschnitten 19.11 und 19.12 sehen, dass man in diesem Modell auch kategoriale unabhängige Variablen sowie gleichzeitig kategoriale unabhängige und metrische unabhängige Variablen untersuchen kann. Zur inferenzstatistischen Absicherung der geschätzten Modellparameter und zur Bestimmung der Konfidenzintervalle werden drei zusätzliche Annahmen getroffen (s. Abschn. 17.9.2):

- (1) Homoskedastizität
- (2) Normalverteilung der Residualvariablen
- (3) Unabhängigkeit der Residuen

Diese Annahmen haben dieselbe Bedeutung, die wir in Abschnitt 17.9 ausführlich behandelt haben. Übertragen auf die multiple Regressionsanalyse bedeutet die Homoskedastizität:

$$\text{Var}(Y | X_1, \dots, X_k) = \text{Var}(\varepsilon | X_1, \dots, X_k) = \sigma_\varepsilon^2 \quad (\text{F 19.28})$$

Die bedingte Normalverteilung bezieht sich ebenfalls auf die Bedingung aller unabhängigen Variablen.

Modell mit stochastischen Regressoren

Im Modell mit einer stochastischen unabhängigen Variablen wird davon ausgegangen, dass es sich bei den Werten x der unabhängigen Variablen X um die Realisierungen einer Zufallsvariablen handelt. In Analogie zum Modell mit einer unabhängigen Variablen wird im Modell der multiplen Regression angenommen, dass alle betrachteten $k + 1$ Variablen multivariat normalverteilt sind. Die multivariate Normalverteilung ist die Erweiterung der bivariaten Normalverteilung auf den Fall von mehr als zwei Variablen. Sie lässt sich grafisch nicht mehr einfach darstellen, und ihre Dichtefunktion lässt sich nur matrixalgebraisch definieren, worauf wir verzichten. Hat man einen Satz von Variablen, die multivariat normalverteilt sind, folgt hieraus, dass jedes Paar von Variablen bivariat normalverteilt und jede einzelne Variable univariat normalverteilt ist. Der Umkehrschluss gilt hingegen nicht notwendigerweise: Liegen k univariat normalverteilte Variablen vor, folgt nicht zwangsläufig, dass die gemeinsame Verteilung dieser Variablen einer k -variaten (k -dimensionalen) Normalverteilung entspricht.

19.7.3 Schätzung der Residualvarianz und des Standardschätzfehlers

Die Schätzung der Residualvarianz folgt demselben Prinzip wie bei der einfachen linearen Regressionsanalyse (s. Abschn. 17.9.3). Die Quadratsumme dieser Residuen wird nun durch $n - k - 1$ geteilt, da man durch die Schätzung von k Regressionsgewichten und einem Achsenabschnitt $k + 1$ Freiheitsgrade verliert:

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= \frac{1}{n - k - 1} \cdot \sum_{m=1}^n e_m^2 \\ &= \frac{1}{n - k - 1} \cdot \sum_{m=1}^n (y_m - \hat{y}_m)^2\end{aligned}\quad (\text{F 19.29})$$

Die Wurzel aus dieser geschätzten Populationsfehlervarianz ist der geschätzte *Standardschätzfehler* der multiplen Regressionsanalyse. Er gibt an, wie stark in der Population die beobachteten y -Werte um die vorhergesagten Werte streuen.

19.7.4 Schätzung, Signifikanztest und Konfidenzintervalle für die multiple Korrelation und den Determinationskoeffizienten

Wie wir bereits in Abschnitt 17.9.8 gesehen haben, ist der Determinationskoeffizient R^2 keine erwartungstreue Schätzung des Populations-Determinationskoeffizienten P^2 (P : großes griechisches Rho). Vielmehr gilt (Wishart, 1931):

$$E(R^2 | P^2 = 0) = \frac{k}{n - 1} \quad (\text{F 19.30})$$

Diese Formel zeigt, dass die Verzerrung umso kleiner wird, je größer n ist. Für kleine Stichproben wurden verschiedene Korrekturformeln entwickelt, über die Carter (1979) einen Überblick und Anwendungsempfehlungen gibt. Wie wir bereits in Kapitel 16 und 17 erwähnt haben, hat sich in verschiedenen Kontexten die Korrekturformel nach Olkin und Pratt (1958) bewährt. Diese lautet für die multiple Regressionsanalyse:

$$\begin{aligned}R_{\text{korrr-OP}}^2 &= 1 - \frac{n - 3}{n - k - 1} \\ &\quad \cdot \left[(1 - R^2) + \frac{2}{n - k + 1} \cdot (1 - R^2)^2 \right]\end{aligned}\quad (\text{F 19.31})$$

Die Bezeichnung $R_{\text{korrr-OP}}^2$ mit OP für Olkin und Pratt haben wir gewählt, weil wir weiter unten noch eine zweite Korrektur von R^2 behandeln werden. Von unterschiedlichen Statistikprogrammen werden unterschiedliche Korrekturformeln herangezogen. Bei großen Stichproben weichen die korrigierten Determinationskoeffizienten nur unwesentlich von den unkorrigierten ab, sodass in diesem Fall auf eine Korrektur verzichtet werden kann.

Signifikanztest

Zunächst interessiert bei der multiplen Regressionsanalyse, ob die unabhängigen Variablen insgesamt einen Beitrag zur Erklärung von Unterschieden in der abhängigen Variablen leisten. Die Nullhypothese lautet, dass die multiple Korrelation in der Population gleich 0 ist:

$$H_0: P = 0$$

Die entsprechende Alternativhypothese hierzu lautet, dass die multiple Korrelation in der Population größer als 0 ist:

$$H_1: P > 0$$

Diese Formulierung ist äquivalent zu der Aussage, dass der Determinationskoeffizient in der Population den Wert 0 annimmt. Sie ist auch äquivalent zu der Aussage, dass der Einfluss aller unabhängigen Variablen nicht bedeutsam ist, d.h., dass in der Population alle Regressionsgewichte β_j den Wert 0 aufweisen:

$$H_0: \beta_1 = \dots = \beta_j = \dots = \beta_k = 0$$

Hierzu korrespondiert die Alternativhypothese, dass mindestens ein Regressionsgewicht β_j in der Population von 0 verschieden ist:

$$H_1: \text{mindestens ein } \beta_j \neq 0$$

Zur Testung der Nullhypothese existiert eine Prüfstatistik, die unter der Nullhypothese, dass der Determinationskoeffizient in der Population gleich 0 ist, einer F -Verteilung mit $df_1 = k$ Zähler- und $df_2 = n - k - 1$ Nennerfreiheitsgraden folgt:

$$F = \frac{n - k - 1}{k} \cdot \frac{R^2}{(1 - R^2)} \quad (\text{F 19.32})$$

Bei dieser Teststatistik wird der Determinationskoeffizient R^2 mit dem Indeterminationskoeffizienten $1 - R^2$ verglichen. Dieses Verhältnis wird multipliziert

mit dem Verhältnis zwischen Nennerfreiheitsgraden ($df_2 = n - k - 1$) und Zählerfreiheitsgraden ($df_1 = k$). Durch Umformen und aufgrund der Formeln F 19.23 und F 19.24 ergibt sich, dass dieser F -Wert gleichbedeutend mit folgendem Quotienten ist:

$$F = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}} = \frac{\frac{\sum_{m=1}^n (\hat{y}_m - \bar{y})^2}{k}}{\frac{\sum_{m=1}^n (y_m - \hat{y}_m)^2}{n - k - 1}} \quad (\text{F 19.33a})$$

Im Zähler dieses F -Werts steht die Quadratsumme der vorhergesagten Werte (QSR) und somit die durch die Regression determinierte Variation geteilt durch die Zählerfreiheitsgrade. Im Nenner steht die Residualquadratsumme (QSE) geteilt durch die Nennerfreiheitsgrade. Teilt man die einzelnen Quadratsummen durch die zugeordneten Freiheitsgrade, so erhält man – wie wir bereits bei der Varianzanalyse (s. Kap. 13) gesehen

haben – die mittleren Quadratsummen. Wie bei der Varianzanalyse wird die F -verteilte Prüfgröße gebildet, indem die mittlere Quadratsumme der Regression (MQSR) durch die mittlere Quadratsumme der Residuen (MQSE) geteilt wird:

$$F = \frac{MQSR}{MQSE} \quad (\text{F 19.33b})$$

Daher lassen sich die Ergebnisse einer multiplen Regressionsanalyse auch in Form einer Varianzanalysetafel darstellen (s. Tab. 19.4). Für unser empirisches Beispiel mit der abhängigen Variablen Wohlbefinden und den beiden unabhängigen Variablen Aufrechterhaltung guter Stimmung und Verbesserung schlechter Stimmung ergibt sich die Varianzanalysetafel in Tabelle 19.5. Man sieht, dass die beiden Prädiktorvariablen zusammen einen bedeutsamen Einfluss auf die abhängige Variable haben, wenn man ein $\alpha = 0,05$ zugrunde legt: Der p -Wert des F -Wertes ist deutlich kleiner als 0,05.

Tabelle 19.4 Darstellung der Ergebnisse einer Regressionsanalyse in Form einer Varianzanalysetafel

Quelle der Variation	Quadratsumme (QS)	Freiheitsgrade (df)	Mittlere Quadratsumme	F-Wert	p-Wert
Regression	QSR	k	MQSR	$\frac{MQSR}{MQSE}$	
Residuen	QSE	$n - k - 1$	MQSE		
Gesamt	QST	$n - 1$	MQST		

Tabelle 19.5 Regression des subjektiven Wohlbefindens auf die beiden Facetten der Stimmungsregulationskompetenz: Ergebnisse in Form einer Varianzanalysetafel sowie geschätzte Determinationskoeffizienten ($n = 237$)

Quelle der Variation	Quadratsumme (QS)	Freiheitsgrade (df)	Mittlere Quadratsumme	F-Wert	p-Wert
Regression	29,661	2	14,831	59,700	< 0,01
Residuen	58,130	234	0,248		
Gesamt	87,791	236	0,372		

$$R^2 = \frac{QSR}{QST} = 1 - \frac{QSE}{QST} = 1 - \frac{58,130}{87,791} = 0,338$$

$$R^2_{\text{korrr-OP}} = 1 - \frac{n-3}{n-k-1} \cdot \left[(1-R^2) + \frac{2}{n-k+1} \cdot (1-R^2)^2 \right] = 1 - \frac{234}{234} \left[0,662 + \frac{2}{236} (1-0,338)^2 \right] = 0,334$$

$$R^2_{\text{korrr-MQS}} = 1 - \frac{MQSE}{MQST} = 1 - \frac{0,248}{0,372} = 0,333$$

Die mittlere Residualquadratsumme von $MQSE = 0,248$ entspricht der geschätzten Fehlervarianz $\hat{\sigma}_\varepsilon^2$. Ihre Quadratwurzel ist der geschätzte Standardschätzfehler $\hat{\sigma}_\varepsilon = 0,498$, der in einigen Statistikprogrammen auch Standardfehler des Schätzers genannt wird.

Teilt man die mittlere Residualquadratsumme ($MQSE$) durch die mittlere gesamte (totale) Quadratsumme ($MQST$), erhält man eine Schätzung des Indeterminationskoeffizienten, auf die in Statistikprogrammen häufig zurückgegriffen wird, da sie weniger verzerrt ist als das Verhältnis der Quadratsummen. Diese Verzerrung ist dadurch bedingt, dass der Determinationskoeffizient R^2 kein erwartungstreuer Schätzer des Populations-Determinationskoeffizienten ist. Hieraus ergibt sich eine weitere korrigierte Schätzung des Determinationskoeffizienten aufgrund der mittleren Quadratsummen (MQS):

$$R_{\text{korrr-MQS}}^2 = 1 - \frac{MQSE}{MQST} \quad (\text{F 19.34})$$

Die Ergebnisse in Tabelle 19.5 zeigen, dass sich das R^2 (0,339) und die beiden korrigierten R^2 -Werte nach F 19.31 (0,334) und F 19.34 (0,333) nur unwesentlich voneinander unterscheiden. Dies liegt an der relativ großen Stichprobe von $n = 237$. Durch die beiden Stimmungsregulationskompetenzen werden ungefähr 33 % der Varianz des Wohlbefindens bestimmt.

Konfidenzintervalle

Wie wir schon bei der einfachen linearen Regressionsanalyse in Abschnitt 17.9.8 gesehen haben, geht man bei der Bestimmung des Konfidenzintervalls des Populations-Determinationskoeffizienten P^2 so vor, dass der geschätzte Determinationskoeffizient R^2 zunächst in den geschätzten Nonzentralitätsparameter einer nonzentralen F -Verteilung übertragen wird. Um diesen geschätzten Nonzentralitätsparameter wird ein Konfidenzintervall für den Populations-Nonzentralitätsparameter gelegt. Die untere und die obere Intervallgrenze werden dann in Intervallschranken für den Determinationskoeffizienten retransformiert, wodurch man ein Konfidenzintervall für den Determinationskoeffizienten erhält. Der Nonzentralitätsparameter der nonzentralen F -Verteilung wird wie folgt geschätzt:

$$\hat{\lambda} = n \cdot \frac{R^2}{1 - R^2} \quad (\text{F 19.35})$$

Um ein $(1 - \alpha)$ -Konfidenzintervall für λ zu schätzen, bestimmt man z.B. mit dem Computerprogramm NDC (Steiger, o. J.) die Intervallgrenzen λ_u und λ_o . Diese können dann mittels der Gleichungen F 17.50 in die unteren und oberen Grenzen des Intervalls für P^2 umgerechnet werden:

$$P_u^2 = \frac{\lambda_u}{\lambda_u + n} \quad \text{und} \quad P_o^2 = \frac{\lambda_o}{\lambda_o + n} \quad (\text{F 19.36})$$

Man kann die Grenzen des Konfidenzintervalls auch direkt mit dem Statistikprogramm R und dem Paket MBESS (Kelley, 2007) berechnen. Aus den bereits in Abschnitt 17.9.8 genannten Gründen wird beim Determinationskoeffizienten häufig ein $(1 - 2 \cdot \alpha)$ -Konfidenzintervall berechnet.

Beispiel

Wohlbefinden und Stimmungsregulationskompetenz: Konfidenzintervall des Populations-Determinationskoeffizienten P^2

Nach dem in Abschnitt 17.9.8 beschriebenen Vorgehen bestimmen wir für unser Anwendungsbeispiel zunächst mit dem Programm NDC (Steiger, o. J. [↓](#)) die Grenzen des zweiseitigen 90 %-Konfidenzintervalls des Nonzentralitätsparameters. Sie betragen $\lambda_u = 81,411$ und $\lambda_o = 161,735$. Um diese Werte berechnen zu lassen, müssen wir dem Programm den F -Wert von 59,700 mitteilen. Die berechneten Intervallgrenzen werden dann nach Gleichung F 19.36 mit $n = 237$ in die Grenzen des Intervalls für P^2 umgerechnet, wodurch man $P_u^2 = 0,256$ und $P_o^2 = 0,406$ und somit das zweiseitige 90 %-Konfidenzintervall $[0,256; 0,406]$ erhält.

Die Bestimmung des in diesem Abschnitt behandelten Konfidenzintervalls setzt deterministische Regressoren voraus. In unserem Beispiel haben wir die Werte der Regressoren jedoch nicht a priori festgelegt, sondern sie als Ergebnis einer Zufallsziehung registriert. Das zugrunde liegende Modell entspricht also dem Modell mit stochastischen Regressoren. In diesem Fall kann der Determinationskoeffizient auch als Schätzer für die quadrierte multiple Korrelation der beiden Prädiktoren mit dem Kriterium in der Population angesehen werden, da ja nun auch die Werte auf den unabhängigen Variablen durch Zufallsauswahl entstanden sind und ihre Verteilungen nicht a priori festgelegt wurden.

Deshalb müssen wir den Stichprobenfehler der Ausprägungen der unabhängigen Variablen berücksichtigen, wodurch sich – wie wir bereits in Abschnitt 17.9.8 gesehen haben – ein größeres Konfidenzintervall ergibt. Entsprechend folgt die Stichprobenkennwerteverteilung des Determinationskoeffizienten R^2 nicht mehr einer nonzentralen F -Verteilung, sondern einer komplexeren Verteilung. Um das Konfidenzintervall zu bestimmen, wendet man das bereits in Abschnitt 17.9.8 beschriebene Vorgehen an. Berechnet man das Konfidenzintervall für das Modell mit einem stochastischen Regressor mit dem Computerprogramm R2 (Steiger und Fouladi, 1997), ergibt sich das zweiseitige 90 %-Konfidenzintervall $[0,252; 0,416]$, das geringfügig größer ausfällt als das Konfidenzintervall für deterministische Prädiktoren.

Konfidenzintervalle lassen sich auch für korrigierte R^2 -Werte bestimmen. Inferenzstatistische Entscheidungen, die man auf der Grundlage dieser Konfidenzintervalle trifft, stehen aber nicht mehr im Einklang mit den entsprechenden Signifikanztests und weisen noch andere Nachteile auf (Smithson, 2003), sodass sie üblicherweise nicht berechnet werden.

A-priori-Poweranalyse: Bestimmung der optimalen Stichprobengröße

Um den optimalen Stichprobenumfang zu bestimmen, folgt man dem bereits in Abschnitt 17.9.8 beschriebenen Vorgehen. Für das Modell mit deterministischen Regressoren legt man den erwarteten Determinationskoeffizienten P_1^2 in der Population, das α - und das β -Niveau vorher fest. Die Berechnung des optimalen Stichprobenumfangs basiert auf der Effektgröße

$$\phi_1^2 = \frac{P_1^2}{1 - P_1^2}. \quad (\text{F 19.37})$$

Zur Bestimmung des optimalen Stichprobenumfangs bestimmt man den Nonzentralitätsparameter einer nonzentralen F -Verteilung, für die gilt, dass derjenige F -Wert, der unter der zentralen Verteilung einen Flächenanteil von α nach rechts abschneidet, unter der nonzentralen Verteilung einen Flächenanteil von β nach links abschneidet. Aus dem Nonzentralitätsparameter dieser Verteilung erhält man dann die optimale Stichprobengröße nach der Formel

$$n = \lambda_1 \cdot \frac{1 - P_1^2}{P_1^2}. \quad (\text{F 19.38})$$

Zusätzlich zum Nonzentralitätsparameter muss nun noch der Populationseffekt, also der Populations-Determinationskoeffizient P_1^2 , spezifiziert werden. Der Nonzentralitätsparameter λ_1 sowie der dazugehörige optimale Stichprobenumfang können z.B. mit dem Computerprogramm G*Power (Faul et al., 2007) bestimmt werden.

Im Falle stochastischer Regressoren legt man ebenfalls den Determinationskoeffizienten, das α - und das β -Niveau vorher fest und bestimmt dann die optimale Stichprobengröße mithilfe eines Computerprogramms wie z.B. R2 von Steiger und Fouladi (1997), das die Stichprobenkennwerteverteilung von R^2 unter der Alternativhypothese adäquat bestimmt.

Beispiel

Bestimmung des optimalen Stichprobenumfangs

Angenommen, man will anhand zweier Prädiktoren 25 % der Varianz einer Kriteriumsvariablen aufklären. Man trifft folgende Festlegungen: $\alpha = 0,05$; $\beta = 0,20$; $P_1^2 = 0,25$. Dann benötigt man nach G*Power $n = 33$ Personen für das Modell mit deterministischen Regressoren und nach R2 $n = 35$ Personen für das Modell mit stochastischen Prädiktoren. Der Unterschied zwischen beiden Modellen ist also relativ gering.

19.7.5 Schätzung, Signifikanztest und Konfidenzintervalle für einen Partialregressionskoeffizienten β_j

Muss die Nullhypothese, dass die multiple Korrelation gleich 0 ist, verworfen werden, stellt sich als Nächstes die Frage, welche der unabhängigen Variablen einen bedeutsamen Einfluss auf das Kriterium haben. Hierzu kann der Einfluss einzelner unabhängiger Variablen überprüft werden. Dabei wird die Nullhypothese geprüft, dass ein einzelner Partialregressionskoeffizient β_j ($j = 1, \dots, k$) in der Population gleich 0 ist:

$$H_0: \beta_j = 0$$

Dem steht die Alternativhypothese gegenüber, dass der Partialregressionskoeffizient ungleich 0 ist:

$$H_1: \beta_j \neq 0$$

Diese beiden Hypothesen sind ein Spezialfall der allgemeineren Hypothese, dass der Regressionskoeffizient β_j in der Population einen spezifischen Wert β_{0j} annimmt:

$$H_0: \beta_j = \beta_{0j} \quad \text{und} \quad H_1: \beta_j \neq \beta_{0j}$$

Gerichtete Hypothesen lassen sich entsprechend formulieren.

Zur Schätzung des Regressionsgewichts β_j in der Population greift man auf das Regressionsgewicht b_j zurück, das man in einer Stichprobe anhand der Kleinst-Quadrate-Schätzung bestimmen kann. Dieses stellt eine erwartungstreue Schätzung des Regressionsgewichts β_j in der Population dar. Der Standardfehler des Regressionsgewichts lässt sich in der multiplen Regressionsanalyse wie folgt bestimmen:

$$\hat{\sigma}_{B_j} = \frac{s_Y}{s_{X_j}} \sqrt{\frac{1 - R^2}{n - k - 1}} \cdot \sqrt{\frac{1}{1 - R_{X_j|X_{\setminus j}}^2}} \quad (\text{F 19.39})$$

Zur Erinnerung: Der Index des Standardfehlers ist ein großes B , da der Standardfehler die Standardabweichung der Stichprobenkennwertverteilung ist. Ein Wert b_j von B_j ist die konkrete Realisierung der Zufallsvariablen B_j in einer Untersuchung. Der Ausdruck $R_{X_j|X_{\setminus j}}^2$ bezeichnet die quadrierte multiple Korrelation der unabhängigen Variablen X_j mit allen anderen unabhängigen Variablen. Ist diese quadrierte multiple Korrelation besonders groß, so wird das Regressionsgewicht ungenau geschätzt; d.h., je stärker eine unabhängige

Variable mit allen anderen unabhängigen Variablen korreliert, umso ungenauer ist die Bestimmung ihres Regressionsgewichts anhand von Stichprobendaten. Wir werden auf dieses sog. Multikollinearitätsproblem in Abschnitt 19.13.4 zurückkommen.

Anhand des Standardfehlers in F 19.39 kann die Hypothese $H_0: \beta_j = \beta_{0j}$ überprüft werden. Hierzu teilt man die Abweichung $(b_j - \beta_{0j})$ durch den geschätzten Standardfehler des Regressionsgewichts und erhält folgende t -verteilte Prüfgröße:

$$T = \frac{B_j - \beta_{0j}}{\hat{\sigma}_{B_j}} \quad (\text{F 19.40})$$

Die Freiheitsgrade der t -Verteilung betragen $df = n - k - 1$. Für eine konkrete Untersuchung mit einem geschätzten Regressionsgewicht b_j erhält man den t -Wert

$$t = \frac{b_j - \beta_{0j}}{\hat{\sigma}_{B_j}}.$$

Will man die Nullhypothese überprüfen, dass die Variable X_j keinen über die anderen unabhängigen Variablen hinausgehenden Prognose- bzw. Erklärungsbeitrag leistet, setzt man den Parameter β_{0j} in Gleichung F 19.40 gleich 0.

Auf der Grundlage des geschätzten Standardfehlers eines Regressionsgewichts lässt sich ein $(1 - \alpha)$ -Konfidenzintervall für das Regressionsgewicht bestimmen:

$$b_j \pm \hat{\sigma}_{B_j} \cdot t_{(1-\frac{\alpha}{2}; df)} \quad (\text{F 19.41})$$

Beispiel

Wohlbefinden und Stimmungsregulationskompetenz: Regressionsgewichte

In Tabelle 19.6 sind die Ergebnisse für unser Beispiel tabellarisch zusammengestellt. Die Ergebnisse zeigen, dass beide Regulationskompetenzen einen bedeutsamen Einfluss auf das Wohlbefinden haben. Die beiden Konfidenzintervalle enthalten nicht den Wert 0, was bedeutet, dass die beiden geschätzten Regressionsgewichte von 0 verschieden sind.

Will man die gerichtete Alternativhypothese testen, dass die Partialregressionskoeffizienten in der Population größer als 0 sind, müssen die p -Werte

in Tabelle 19.6 halbiert werden. Da in unserem Beispiel schon der zweiseitige Test zu einem signifikanten Ergebnis führt, muss dies zwangsläufig auch für den einseitigen Test gelten. Die untere Grenze des einseitigen Konfidenzintervalls enthält man in diesem Fall beispielsweise für das Regressionsgewicht β_1 mit $b_1 - \hat{\sigma}_{B_1} \cdot t_{(1-\alpha; n-k-1)} = 0,354 - \hat{\sigma}_{B_1} \cdot t_{(0,95; 234)} = 0,354 - 0,06 \cdot 1,651 = 0,255$. Das einseitige 95 %-Konfidenzintervall lautet dann $[0,255; \infty)$.

Tabelle 19.6 Ergebnisse einer multiplen Regressionsanalyse mit der abhängigen Variablen Wohlbefinden und den beiden Stimmungsregulationskompetenzen als unabhängigen Variablen (UV): Signifikanztests und Konfidenzintervalle für die Partialregressionskoeffizienten

UV	b_j	$\hat{\sigma}_{b_j}$	t-Wert	p-Wert	95 %-Konfidenzintervall	
					Untergrenze	Obergrenze
Aufrechterhaltung guter Stimmung	0,354	0,060	5,912	< 0,001	0,236	0,471
Verbesserung schlechter Stimmung	0,336	0,060	5,598	< 0,001	0,217	0,454

Überprüft man einzelne Regressionsgewichte auf ihre Signifikanz, stellt sich die Frage, ob das α -Niveau in Bezug auf die multiplen Tests korrigiert werden muss. Wir haben diesen Sachverhalt und entsprechende Korrekturen ausführlich bei der einfaktoriellen Varianzanalyse in Abschnitt 13.1.11 behandelt. So lässt sich bei der multiplen Regressionsanalyse z. B. auch eine Bonferroni-Korrektur durchführen (Mundfrom et al., 2006). Diese ist jedoch sehr konservativ und führt dazu, dass das hiermit erreichte α -Niveau geringer ist als das festgelegte α -Niveau (Mundfrom et al., 2006). Hierdurch verringert sich auch die Power, Effekte aufzudecken. Cohen et al. (2003) setzen sich aus der Perspektive der Optimierung der Power und adäquaten Absicherung des α -Niveaus intensiv mit der Frage auseinander, ob eine Korrektur des α -Niveaus vorgenommen werden soll oder nicht. Sie empfehlen der von R. A. Fisher vorgeschlagenen Vorgehensweise zu folgen und das α -Niveau der Einzeltests nicht zu korrigieren. Dies setzt aber voraus, dass der globale Test zuvor zu einem signifikanten Ergebnis geführt hat. Darüber hinaus raten sie, die Anzahl der unabhängigen Variablen so klein wie möglich zu halten und nur diejenigen Variablen in die Analyse aufzunehmen, die aufgrund theoretischer Überlegungen relevant sind.

■ Überprüfung der Nützlichkeit eines Prädiktors

Eine zweite, äquivalente Möglichkeit, die Nullhypothese zu überprüfen, dass das Regressionsgewicht einer unabhängigen Variablen in der Population gleich 0 ist, besteht darin, anhand eines F -Tests den Verlust an aufgeklärter Varianz zu überprüfen, wenn die betreffende unabhängige Variable aus der Regressionsgleichung herausgenommen wird. Der Test basiert auf dem Vergleich der Determinationskoeffizienten zweier Regressionsmodelle. Der Determinationskoeffizient des

Modells, das alle unabhängigen Variablen enthält (R_2^2), wird verglichen mit dem Determinationskoeffizienten des Modells, das alle unabhängigen Variablen bis auf die betreffende unabhängige Variable X_j enthält (R_1^2). Der F -Test lautet wie folgt:

$$F = (n - k - 1) \cdot \frac{R_2^2 - R_1^2}{(1 - R_2^2)} \quad (F\ 19.42)$$

Die F -Verteilung hat $df_1 = 1$ Zähler- und $df_2 = n - k - 1$ Nennerfreiheitsgrade. Dass $df_1 = 1$ ist, lässt sich leicht erklären, denn die beiden verglichenen Regressionsmodelle unterscheiden sich lediglich in einem Modellparameter. Der F -Test überprüft, ob die Nützlichkeit der Prädiktorvariablen X_j von 0 verschieden ist. In Tabelle 19.7 wird dieser F -Test für unser Beispiel mit der abhängigen Variablen Wohlbefinden und den beiden Stimmungsregulationskompetenzen als unabhängigen Variablen veranschaulicht. Es wird der Erklärungszuwachs durch die Verbesserung schlechter Stimmung überprüft.

Die Änderung des Determinationskoeffizienten zeigt an, dass durch die Hinzunahme der Verbesserung schlechter Stimmung 8,9 % mehr Varianz der abhängigen Variablen erklärt wird als mit der Aufrechterhaltung guter Stimmung alleine. Dieser Zugewinn ist statistisch bedeutsam von 0 verschieden, da der p -Wert des F -Wertes kleiner als 0,05 ist.

Würde man die exakten Ergebnisse mit mehr als zwei Nachkommastellen angeben, würde man erkennen, dass der F -Wert in Tabelle 19.7 exakt dem quadrierten t -Wert des Regressionsgewichts b_2 in Tabelle 19.6 entspricht. Die Ergebnisse in Tabelle 19.7 wurden mit einem Computerprogramm berechnet. Der F -Wert unterscheidet sich daher von dem F -Wert, den man erhalten würde, wenn man die auf zwei Nachkommastellen

Tabelle 19.7 Vergleich zweier Regressionsmodelle zur Vorhersage des habituellen Wohlbefindens: Überprüfung der Nützlichkeit der Kompetenz, schlechte Stimmung zu verbessern

Modell (UV)	R	R ²	Änderungsstatistiken				
			Änderung in R ²	F-Wert	df ₁	df ₂	p-Wert
Aufrechterhaltung guter Stimmung	0,499	0,249					
Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung	0,581	0,338	0,089	31,342	1	234	< 0,001

gerundeten R^2 -Werte in Gleichung F 19.42 einsetzen würde. Der Unterschied ist aber nur durch den Rundungsfehler bedingt.

Der dargestellte F -Test ist der Spezialfall eines allgemeineren F -Tests zur Überprüfung der Hypothese, dass ein Satz von unabhängigen Variablen keinen bedeutsamen Beitrag zur Varianzaufklärung in der Population liefert, der über einen Satz von anderen unabhängigen Variablen hinausgeht. Diesen Test stellen wir im nächsten Abschnitt vor und zeigen, wie man Konfidenzintervalle und die optimalen Stichprobengröße berechnen kann.

19.7.6 Schätzung, Signifikanztest und Konfidenzintervalle für einen Satz unabhängiger Variablen

Neben der Frage, ob einzelne unabhängige Variablen einen bedeutsamen Einfluss auf die abhängige Variable haben, kann auch die Frage untersucht werden, ob eine Menge von unabhängigen Variablen einen bedeutsamen Einfluss auf die abhängige Variable hat. Hierzu betrachtet man die Determinationskoeffizienten zweier Modelle. In dem sog. uneingeschränkten Modell wird der Einfluss aller unabhängigen Variablen berücksichtigt. Der Determinationskoeffizient in diesem Modell wird mit R_u^2 bezeichnet. Das sog. eingeschränkte Modell enthält alle unabhängigen Variablen bis auf die Menge von unabhängigen Variablen, deren gemeinsamer Einfluss überprüft werden soll. Der Determinationskoeffizient in diesem eingeschränkten Modell wird mit R_e^2 bezeichnet. Die ungerichtete Nullhypothese, die hierbei überprüft wird, lautet, dass alle Partialregressionskoeffizienten der unabhängigen Variablen, deren Einfluss geprüft werden soll, gleich 0 sind. Die dazugehörige Alternativhypothese lautet, dass mindestens einer dieser Partialregressionskoeffizienten ungleich 0 ist.

Wäre man z.B. daran interessiert, ob die beiden Stimmungsregulationskompetenzen einen bedeutsamen Einfluss auf das Wohlbefinden haben, der über die Extraversion hinausgeht, würde man folgende eingeschränkte und uneingeschränkte Modelle betrachten:

- Das uneingeschränkte Modell würde die drei unabhängigen Variablen Extraversion, Verbesserung schlechter Stimmung und Aufrechterhaltung guter Stimmung umfassen.
- Das eingeschränkte Modell würde nur die Extraversion enthalten.

Über den Vergleich der Determinationskoeffizienten des uneingeschränkten und des eingeschränkten Modells prüft man die Nullhypothese $H_0: \beta_2 = \beta_3 = 0$, wobei X_2 die Verbesserung schlechter Stimmung und X_3 die Aufrechterhaltung guter Stimmung erfassen würde. Diese Nullhypothese ist äquivalent zur Nullhypothese $H_0: P_u^2 - P_e^2 = 0$. Zur Überprüfung dieser Nullhypothese gibt es einen F -Test, der wie folgt lautet:

$$F = \frac{(n - k_u - 1)}{k_u - k_e} \cdot \frac{R_u^2 - R_e^2}{(1 - R_u^2)} \quad (\text{F 19.43})$$

In dieser Formel bezeichnen R_u^2 den Determinationskoeffizienten des uneingeschränkten Modells, R_e^2 den Determinationskoeffizienten des eingeschränkten Modells, k_u die Anzahl der unabhängigen Variablen im uneingeschränkten Modell und k_e die Anzahl der unabhängigen Variablen im eingeschränkten Modell. Man beachte die Ähnlichkeit mit Formel F 19.42. Der Unterschied zwischen den beiden Gleichungen besteht lediglich im Nenner des vorderen Quotienten auf der rechten Seite der Gleichung. In Gleichung F 19.42 ist dieser Nenner immer 1, da sich die beiden verglichenen Regressionsmodelle lediglich in einem Modellparameter unterscheiden. Der Nenner erscheint daher gar nicht in

Tabelle 19.8 Vergleich zweier Regressionsmodelle zur Vorhersage des habituellen Wohlbefindens

Modell (UV)	R	R ²	Änderungsstatistiken				
			Änderung in R ²	F-Wert	df ₁	df ₂	p-Wert
Eingeschränkt: Extraversion	0,251	0,063					
Uneingeschränkt: Extraversion; Verbesserung schlechter Stimmung; Aufrechterhaltung guter Stimmung	0,583	0,339	0,276	48,772	2	233	< 0,001

der Gleichung. Die F -Verteilung hat $df_1 = k_u - k_e$ Zähler- und $df_2 = n - k_u - 1$ Nennerfreiheitsgrade.

Der Ausdruck $R_u^2 - R_e^2$ ist die Semipartialdetermination (quadrierte multiple semipartielle Korrelation), die wir in Abschnitt 19.6 ausführlich behandelt haben. Die Ergebnisse für unser Beispiel sind in Tabelle 19.8 zusammengefasst. Sie zeigen, dass die beiden Stimmungsregulationskompetenzen über die Extraversion hinaus bedeutsame Unterschiede im Wohlbefinden erklären. Der zusätzliche Anteil an erklärter Varianz beträgt 27,6 %. Dieser Beitrag ist bei einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ signifikant, da der p -Wert des zugehörigen F -Tests kleiner als 0,05 ist.

Konfidenzintervalle

In Analogie zum Konfidenzintervall für den allgemeinen Determinationskoeffizienten P^2 bestimmt man das Konfidenzintervall zunächst nicht für den Semipartialdeterminationskoeffizienten, sondern greift auf die folgende Effektgröße zurück:

$$\phi_2^2 = \frac{P_u^2 - P_e^2}{1 - P_u^2} \quad (\text{F 19.44a})$$

Diese Effektgröße entspricht der Semipartialdetermination geteilt durch den Indeterminationskoeffizienten des uneingeschränkten Modells und ist daher schwierig zu interpretieren, da sie auch größer als 1 werden kann (etwa wenn $P_u^2 = 0,80$ und $P_e^2 = 0,50$). Es lässt sich allerdings zeigen, dass die Effektgröße ϕ_2^2 sich auch aus dem quadrierten multiplen partiellen Korrelationskoeffizienten $P_{YX_{u \cdot e} \cdot X_e}^2$ herleiten lässt:

$$\phi_2^2 = \frac{P_{YX_{u \cdot e} \cdot X_e}^2}{1 - P_{YX_{u \cdot e} \cdot X_e}^2} \quad (\text{F 19.44b})$$

Die multiple partielle Korrelation $P_{YX_{u \cdot e} \cdot X_e}$ ist die Korrelation zwischen der abhängigen Variablen Y und den unabhängigen Variablen, die im uneingeschränkten

Modell, nicht aber im eingeschränkten Modell enthalten sind, nachdem der Einfluss aller unabhängigen Variablen, die im eingeschränkten Modell enthalten sind, auspartialisiert wurde. Sie bestimmt sich wie folgt:

$$P_{YX_{u \cdot e} \cdot X_e} = \sqrt{\frac{P_u^2 - P_e^2}{1 - P_e^2}} \quad (\text{F 19.45})$$

Der Vorteil der Effektgröße in Gleichung F 19.44b besteht darin, dass sie nur von der quadrierten multiplen partiellen Korrelation abhängt und hierauf aufbauend sehr einfach ein Konfidenzintervall für die quadrierte multiple partielle Korrelation bestimmt werden kann. Man bestimmt daher das Konfidenzintervall für die quadrierte Partialkorrelation und nicht für die quadrierte Semipartialkorrelation, da dieses komplexer zu bestimmen wäre und die Partialkorrelation auch für viele Anwendungsfälle diejenige Information enthält, die von Interesse ist. Dass man für die quadrierte partielle multiple Korrelation relativ einfach ein Konfidenzintervall erhalten kann, sieht man an der Analogie der Struktur von Gleichung F 19.44b zur Gleichung F 19.37.

Ist die multiple partielle Korrelation in der Population von 0 verschieden, dann ist der folgende F -Wert ein Wert einer nonzentralen F -Verteilung:

$$\begin{aligned} F &= \frac{n - k_u - 1}{k_u - k_e} \cdot \frac{R_{YX_{u \cdot e} \cdot X_e}^2}{1 - R_{YX_{u \cdot e} \cdot X_e}^2} \\ &= \frac{n - k_u - 1}{k_u - k_e} \cdot \frac{\frac{R_u^2 - R_e^2}{1 - R_e^2}}{1 - \frac{R_u^2 - R_e^2}{1 - R_e^2}} \end{aligned} \quad (\text{F 19.46})$$

Diese nonzentrale F -Verteilung hat $df_1 = k_u - k_e$ und $df_2 = n - k_u - 1$ Freiheitsgrade und einen Nonzentralitätsparameter

$$\lambda = n \cdot \frac{P_{YX_{u-e} \cdot X_e}^2}{1 - P_{YX_{u-e} \cdot X_e}^2} = n \cdot \phi_2^2. \quad (\text{F 19.47})$$

Um ein Konfidenzintervall für $P_{YX_{u-e} \cdot X_e}^2$ zu erhalten, bestimmt man zunächst die Grenzen des Intervalls für den Nonzentralitätsparameter, das man auf der Grundlage des geschätzten Nonzentralitätsparameters bestimmt, den man wie folgt erhält:

$$\hat{\lambda} = n \cdot \frac{R_{YX_{u-e} \cdot X_e}^2}{1 - R_{YX_{u-e} \cdot X_e}^2} \quad (\text{F 19.48})$$

Die Grenzen des $(1 - \alpha)$ -Konfidenzintervalls lassen sich bestimmen, indem die Nonzentralitätsparameter der nonzentralen F -Verteilungen bestimmt werden, von denen der Ausdruck F 19.46 jeweils $\alpha/2$ Flächenanteile am oberen (λ_u) bzw. unteren Ende (λ_o) abschneidet. Diese Schranken des Konfidenzintervalls können z. B. mit dem Programm NDC bestimmt werden. Diese Intervallschranken lassen sich dann in Intervallschranken für die quadrierte partielle multiple Korrelation wie folgt umrechnen:

$$\text{unteres } P_{YX_{u-e} \cdot X_e}^2 = \frac{\lambda_u}{\lambda_u + n}$$

und

$$\text{oberes } P_{YX_{u-e} \cdot X_e}^2 = \frac{\lambda_o}{\lambda_o + n} \quad (\text{F 19.49})$$

Auch für $P_{YX_{u-e} \cdot X_e}^2$ berichtet man häufig das 90 %-Konfidenzintervall, da die quadrierte partielle multiple Korrelation eine untere Schranke von 0 aufweist. Das 90 %-Konfidenzintervall kann dann zur Signifikanztestung auf einem α -Niveau von $\alpha = 0,05$ herangezogen werden (s. hierzu die Ausführungen in Abschn. 17.9.8).

Im Falle stochastischer Regressoren ist uns kein vergleichbarer Ansatz bekannt.

Beispiel

Wohlbefinden, Stimmungsregulationskompetenz und Extraversion: Konfidenzintervall

Für unser Beispiel bestimmen wir zunächst anhand von Tabelle 19.8 den F -Wert gemäß Gleichung F 19.46:

$$F = \frac{n - k_u - 1}{k_u - k_e} \cdot \frac{\frac{R_u^2 - R_e^2}{1 - R_e^2}}{1 - \frac{R_u^2 - R_e^2}{1 - R_e^2}}$$

$$\begin{aligned} &= \frac{237 - 3 - 1}{3 - 1} \cdot \frac{\frac{0,339 - 0,063}{1 - 0,063}}{1 - \frac{0,339 - 0,063}{1 - 0,063}} \\ &= 116,5 \cdot \frac{0,295}{0,705} = 48,748 \end{aligned}$$

Dieser F -Wert hat $df_1 = 2$ Zähler- und $df_2 = 233$ Nennerfreiheitsgrade. Wir bestimmen nun mithilfe des Programms NDC (Steiger, 2004; [↓](#)) für den gefundenen F -Wert von 48,748 als untere Grenze des zweiseitigen 90 %-Konfidenzintervalls den Nonzentralitätsparameter $\lambda_u = 64,034$ und als obere Grenze den Nonzentralitätsparameter $\lambda_o = 135,276$. Diese Intervallgrenzen werden dann nach Gleichung F 19.49 mit $n = 237$ in die unteren und oberen Grenzen des 90 %-Konfidenzintervalls für die quadrierte partielle multiple Korrelation umgerechnet, wodurch man das Konfidenzintervall $[0,213; 0,363]$ erhält. Das 90 %-Konfidenzintervall für die Effektstärke ϕ_2^2 lautet übrigens unter Rückgriff auf Gleichung F 19.44b $[0,271; 0,570]$.

A-priori-Poweranalyse: Bestimmung der optimalen Stichprobengröße

Um den optimalen Stichprobenumfang zu bestimmen, legt man die quadrierte multiple partielle Korrelation in der Population, das α - und das β -Niveau vorher fest. Dazu bestimmt man den Nonzentralitätsparameter einer nonzentralen F -Verteilung, für die gilt, dass derjenige F -Wert, der unter der zentralen F -Verteilung einen Flächenanteil von α nach rechts abschneidet, unter der nonzentralen F -Verteilung einen Flächenanteil von β nach links abschneidet. Der Nonzentralitätsparameter dieser nonzentralen F -Verteilung sowie der dazugehörige optimale Stichprobenumfang kann z. B. mit dem Computerprogramm G*Power (Faul et al., 2007) bestimmt werden. Angenommen, man erwartet für zwei Prädiktoren eine quadrierte partielle multiple Korrelation von 0,25 nach Auspartialisieren einer Variablen. Man trifft folgende Festlegungen: $\alpha = 0,05$; $\beta = 0,20$; $P_{YX_{u-e} \cdot X_e}^2 = 0,25$. Dann benötigt man nach G*Power $n = 33$ Personen. Im Falle stochastischer Regressoren ist uns kein vergleichbares Vorgehen bekannt.

19.7.7 Verfahren zur Auswahl unabhängiger Variablen

Die in den letzten beiden Abschnitten dargestellten Sachverhalte kann man sich zunutze machen, um unabhängige Variablen für ein Regressionsmodell auszuwählen. Hierbei lassen sich zwei generelle Strategien unterscheiden:

- (1) die Auswahl von unabhängigen Variablen aufgrund theoretischer Überlegungen und
- (2) die datengesteuerte Auswahl von Variablen zur Maximierung der Varianzaufklärung der abhängigen Variablen bei gleichzeitiger Minimierung der Anzahl zu berücksichtigender unabhängiger Variablen.

Theoretische Auswahl

Bei der theoretischen Auswahl nimmt man diejenigen unabhängigen Variablen in eine Regressionsanalyse auf, von denen man aufgrund theoretischer Überlegungen eine Varianzaufklärung erwartet. In verschiedenen Forschungskontexten ist es sinnvoll, die unabhängigen Variablen zu Gruppen zusammenzufassen und diese nacheinander blockweise in die Regressionsanalyse aufzunehmen. Hierdurch kann man spezifische Hypothesen überprüfen. In unserem Beispiel könnte man daran interessiert sein, Unterschiede im Wohlbefinden auf Unterschiede in der Stimmungsregulationskompetenz und Unterschiede in der Persönlichkeit zurückzuführen. Die Stimmungsregulationskompetenz setzt sich aus mehreren Komponenten oder Facetten zusammen (z.B. Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung), die Persönlichkeit aus mehreren Eigenschaften (z.B. Extraversion, Neurotizismus, Gewissenhaftigkeit etc.). Bei einer blockweisen Aufnahme der unabhängigen Variablen würde man alle Variablen, die Regulationskompetenzen erfassen, sowie alle Persönlichkeitsvariablen zu jeweils einem Block zusammenfassen. Aus theoretischer Sicht würde man vielleicht zunächst den Block mit den Stimmungsregulationskompetenzen aufnehmen, da man diesen eine größere Rolle für die Vorhersage des Wohlbefindens zuschreibt als den Persönlichkeitseigenschaften. In einem zweiten Schritt würde man den Block mit Persönlichkeitsvariablen aufnehmen, um zu prüfen, ob Persönlichkeitsmerkmale einen inkrementellen Beitrag zur Aufklärung der Varianz des Wohlbefindens leisten. Unter einem inkrementellen Beitrag versteht man den-

jenigen Beitrag, den die neu aufgenommenen Variablen zur Varianzaufklärung leisten, der über die bereits im Modell enthaltenen Variablen hinausgeht.

Die Aufnahme-strategien können aber auch anders gestaltet werden. So empfehlen z.B. Cohen et al. (2003), die Variablen nach kausaler Priorität aufzunehmen. Hierbei würde man zunächst die angenommene Ursache als erste unabhängige Variable aufnehmen und dann schrittweise diejenigen Variablen aufnehmen, die als intervenierende Variablen (s. Abschn. 4.1.3) den Effekt der angenommenen Ursache auf die abhängige Variablen vermitteln könnten. Hierdurch kann man herausfinden, wie groß der Anteil des Effektes der angenommenen Ursache ist, der über die intervenierenden Variablen vermittelt wird, und wie groß der Anteil der angenommenen Ursache an der Varianzaufklärung der abhängigen Variable ist, der nicht vermittelt wird. So könnte man in einer randomisierten Interventions-Kontrollgruppen-Studie beispielsweise daran interessiert sein zu überprüfen, wie groß der Effekt einer Dankbarkeitsintervention auf das Wohlbefinden ist. Als intervenierende Variable könnte man erfassen, wie häufig die Untersuchungsteilnehmer im Alltag Dankbarkeit gezeigt haben (sowohl in der Interventions- als auch der Kontrollgruppe). In der Regressionsanalyse würde man dann zunächst die dichotome Variable Intervention (mit den Ausprägungen Interventions- und Kontrollgruppe) aufnehmen, um zu sehen, ob die Intervention überhaupt einen bedeutsamen Effekt hatte (wir werden in Abschnitt 19.11 sehen, dass wir in die Regressionsanalyse auch dichotome unabhängige Variablen aufnehmen können). In einem nächsten Schritt könnte man die Dankbarkeitsvariable aufnehmen und überprüfen, wie viel Varianz des Wohlbefindens die Intervention über das Dankbarkeitsverhalten hinaus erklärt. Wir werden in Abschnitt 26.1 noch weitere Strategien kennenlernen, um solche Mediatorhypothesen gezielter zu überprüfen.

Die Aufnahmereihenfolge kann aber auch unter pragmatischen Gesichtspunkten erfolgen. Beispielsweise kann es in der Prognoseforschung Variablen geben, die kostengünstiger zu erfassen sind als andere. In einem solchen Fall möchte man wissen, ob die teureren Variablen über die kostengünstigeren Variablen hinaus noch gebraucht werden, um die abhängige Variable vorherzusagen. So ist zur Prognose des Berufserfolgs die Schulnote sehr kostengünstig zu erfassen, den Intelligenzquotienten zu diagnostizieren ist hingegen

vergleichsweise teuer. Deshalb könnte man zunächst die Schulnote in die Regressionsanalyse zur Vorhersage des Berufserfolgs aufnehmen, um dann zu überprüfen, ob die Intelligenz einen inkrementellen Vorhersagebeitrag leistet. Es gibt somit verschiedene Strategien, die unabhängigen Variablen zu gruppieren und in die Regressionsanalyse aufzunehmen. Für welche man sich entscheidet, hängt von der Fragestellung ab, die mit der Analyse geklärt werden soll.

Datengesteuerte Auswahl

Neben der theoretischen Festlegung von unabhängigen Variablen kann deren Auswahl auch datengesteuert erfolgen. Der datengesteuerten Strategie bedient man sich, wenn keine theoretischen Überlegungen zur kausalen Ordnung der Prädiktoren oder ihrer Relevanz angebracht erscheinen (Cohen et al., 2003). Bei der datengesteuerten Auswahl will man aus einer Menge verfügbarer Prädiktoren diejenigen auswählen, die die Kriteriumsvariable optimal vorhersagen. Alle ausgewählten Variablen sollten einen signifikanten Beitrag zur Vorhersage des Kriteriums leisten. Gleichzeitig sollten Prädiktoren, die keinen signifikanten Beitrag zur Vorhersage des Kriteriums leisten, nicht in das Modell mit aufgenommen werden. Zur datengesteuerten Auswahl gibt es drei Strategien:

- (1) die Vorwärtsselektion,
- (2) die Rückwärtselimination und
- (3) die schrittweise Regression.

Bei allen drei Verfahren muss vorher ein α -Niveau festgelegt werden, das angibt, ab wann eine unabhängige Variable einen signifikanten Beitrag zur Aufklärung der Varianz des Kriteriums liefert. Die Selektion der Variablen auf der Basis des gewählten Signifikanzniveaus erledigt dann das jeweilige Statistikprogramm.

Vorwärtsselektion. Bei der Vorwärtsselektion gibt man dem Statistikprogramm alle unabhängigen Variablen bekannt, die für die Vorhersage des Kriteriums infrage kommen. Das Programm nimmt im ersten Schritt diejenige Variable auf, die am höchsten mit der Kriteriumsvariablen korreliert ist. Im nächsten Schritt nimmt es diejenige Variable auf, deren F -Wert nach Formel F 19.43 am größten und gleichzeitig signifikant ist. Im dritten Schritt wird diejenige Variable aufgenommen, die über die beiden bereits in der Gleichung enthaltenen Variablen hinaus am meisten zusätzliche Varianz

aufklärt, d. h. diejenige der verbliebenen Variablen, deren F -Wert nach Formel F 19.43 am größten und gleichzeitig signifikant ist. Das Verfahren wird abgebrochen, wenn keine der verbliebenen Variablen mehr einen signifikanten zusätzlichen Erklärungsbeitrag leistet. Für unser Beispiel, in dem Wohlbefindensunterschiede auf Unterschiede in den beiden Stimmungsregulationskompetenzen und in der Extraversion zurückgeführt werden, folgt dieses Aufnahmeschema den Ergebnissen in Tabelle 19.1. Zuerst wird die Aufrechterhaltung guter Stimmung aufgenommen, dann die Fähigkeit zur Verbesserung schlechter Stimmung. Danach stoppt das Verfahren. Die Extraversion wird nicht aufgenommen, da sie keinen signifikanten zusätzlichen Beitrag zur Vorhersage des Wohlbefindens erbringt.

Rückwärtselimination. Bei diesem Verfahren geht das Programm umgekehrt vor. Zunächst werden alle unabhängigen Variablen aufgenommen. Dann wird in einem ersten Schritt diejenige Variable entfernt, die den geringsten und einen nicht-signifikanten F -Wert nach Formel F 19.42 aufweist. Im nächsten Schritt wird diejenige Variable eliminiert, die von den in der Regressionsgleichung verbliebenen unabhängigen Variablen den geringsten und einen nicht-signifikanten F -Wert aufweist. Dieses Verfahren wird so lange fortgesetzt, bis es keine unabhängige Variable in der Gleichung mehr gibt, deren F -Wert nicht signifikant ist. Nach diesem Verfahren wird in unserem Beispiel als Erstes die Extraversion entfernt. Alle weiteren Variablen verbleiben in der Gleichung. Manche Statistikprogramme verwenden für den Ausschluss von Prädiktorvariablen bei der Rückwärtselimination ein liberaleres Signifikanzkriterium (z. B. $\alpha = 10\%$) als für den Einschluss bei der Vorwärtsselektion.

Schrittweise Regression. Bei der schrittweisen Regression werden beide Strategien kombiniert. Man startet mit einer Vorwärtsselektion auf der Basis eines vorher festgelegten Signifikanzniveaus für den Einschluss (α_E). Nimmt man sukzessive neue unabhängige Variablen auf, kann es passieren, dass bei einer bestimmten Kombination unabhängiger Variablen der Vorhersagebeitrag einer bereits aufgenommenen Variablen nicht länger signifikant ist. Überschreitet der p -Wert einer solchen Variablen ein vorher festgelegtes Signifikanzniveau für den Ausschluss (α_A), so wird diese Variable entfernt, bevor eine weitere neue Variable aufgenommen wird. In manchen Statistikprogrammen wird das Signifi-

kanzniveau für den Ausschluss von Variablen liberaler angesetzt als das Signifikanzniveau für den Einschluss; die entsprechenden Voreinstellungen in den gängigen Statistikprogrammen lassen sich jedoch leicht ändern.

Prognosegüte und Kreuzvalidierung

Hat man die unabhängigen Variablen für das Regressionsmodell ausgewählt, stellt sich die Frage, wie gut das Modell eine Prognose in zukünftigen Fällen leistet. Die Regressionsparameter wurden anhand einer Stichprobe mit einem bestimmten Stichprobenfehler bestimmt, sodass die Prognosegüte in einer anderen Stichprobe nicht automatisch gleich gut sein muss. Wie kann man diese untersuchen? Zum einen könnte man zusätzliche Stichproben ziehen, um die Prognosegüte in ihnen zu untersuchen. Den damit verbundenen Aufwand will man jedoch nicht immer auf sich nehmen. Stattdessen kann man alternative Strategien anwenden, die unter dem Begriff der Kreuzvalidierung (engl. cross-validation) zusammengefasst werden. Eine Möglichkeit der Kreuzvalidierung besteht darin, den Datensatz in zwei Hälften zu teilen. Anhand der einen Hälfte des Datensatzes sucht man das am besten passende Regressionsmodell (sog. Trainings-Substichprobe). Dann nimmt man die zweite Hälfte des Datensatzes (sog. Test-Substichprobe) und prognostiziert für alle Personen dieser Test-Substichprobe ihre y -Werte anhand ihrer x -Werte und der Regressionsgewichte, die man in der Trainings-Substichprobe gewonnen hat. Die Korrelation zwischen den so vorhergesagten Werten und ihren beobachteten y -Werten ist ein Maß für die Prognosegüte. Weicht diese stark von der multiplen Regression in der Trainingsstichprobe ab, zeigt dies, dass die Prognosegüte nicht von der Trainings-Substichprobe auf die Test-Substichprobe verallgemeinert werden kann.

Dieses Vorgehen lässt sich jedoch mit Kockelkorn (2000) aus zwei Gründen kritisieren: (1) Bei der Suche nach dem besten Regressionsmodell verzichtet man auf einen Teil der Daten, wodurch die Schätzungenauigkeit (Standardfehler der Regressionsgewichte und der vorhergesagten y -Werte) zunimmt. (2) Die Aufteilung in beide Stichproben ist mit einer gewissen Beliebigkeit behaftet. Aus diesen Gründen wird zum einen empfohlen, die Trainings-Substichprobe deutlich größer zu wählen als die Test-Substichprobe. So schlagen z. B. Tabachnik und Fidell (2007) hierfür eine Aufteilung in 80 % und 20 % vor. Zum anderen kann man so vorgehen, dass man alle möglichen oder zumindest sehr

viele Aufteilungen der Gesamtstichprobe in die beiden Substichproben vornimmt und die Ergebnisse zusammenfasst.

Häufig wählt man eine Extremvariante der Kombination beider Strategien. Als Trainings-Substichprobe wird die Stichprobe angesehen, aus der ein einziges Element entfernt wurde. Man bestimmt dann die Regressionsparameter anhand dieser Stichprobe und prognostiziert den y -Wert für dasjenige Element (in der Psychologie meist ein Individuum), das der Stichprobe vorher entnommen wurde. Dieses Verfahren kann man für alle Stichprobenelemente wiederholen und die Abweichungen aufsummieren oder mitteln.

Prognosefehler PRESS. Bezeichnet man mit $\hat{y}_{m \setminus (m)}$ den vorhergesagten y -Wert von m , den man anhand der Parameterschätzungen in der Stichprobe ohne m (daher: $\setminus(m)$) gewonnen hat, erhält man zwei Gütemaße für die Prognose. Das erste Maß ist der Prognosefehler *PRESS* (engl. prediction error sum of squares):

$$PRESS = \sum_{m=1}^n (y_m - \hat{y}_{m \setminus (m)})^2 \quad (\text{F } 19.50)$$

Dieses Maß ist die Summe der quadrierten Abweichungen der wahren von den vorhergesagten Werten. Je stärker die Werte in den Test-Substichproben (hier jeweils ein einzelnes m) von den erwarteten Werten, die man anhand der Trainings-Substichprobe gewonnen hat, abweichen, desto schlechter ist die Prognose, und umso größer wird dann auch der quadrierte Abweichungswert und somit der *PRESS*-Wert. Ein Nachteil des *PRESS*-Wertes ist, dass er von der Stichprobengröße abhängt.

Kreuzvalidierungsfehler CVE. Um die Abhängigkeit des *PRESS*-Wertes von der Stichprobengröße n aufzuheben, teilt man ihn durch n und enthält den Kreuzvalidierungsfehler *CVE* (engl. cross validation error):

$$CVE = \frac{1}{n} \cdot \sum_{m=1}^n (y_m - \hat{y}_{m \setminus (m)})^2 \quad (\text{F } 19.51)$$

Beispiel

Wohlbefinden, Stimmungsregulationskompetenz und Extraversion: Kreuzvalidierungsfehler

Berechnet man für die drei Regressionsmodelle in Tabelle 19.1 den Kreuzvalidierungsfehler, so erhält man folgende Werte:

- ▶ Modell I (Aufrechterhaltung guter Stimmung): $CVE_I = 0,283$
- ▶ Modell II (Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung): $CVE_{II} = 0,253$
- ▶ Modell III (Aufrechterhaltung guter Stimmung, Verbesserung schlechter Stimmung, Extraversion): $CVE_{III} = 0,254$

Man sieht, dass der Kreuzvalidierungsfehler im zweiten Modell am geringsten ist. Auch dieser Befund spricht dafür, das zweite Modell auszuwählen. Mit Statistikprogrammen kann man sich die »gelöschten« oder »ausgeschlossenen« Residuen ausgeben lassen. Diese muss man dann quadrieren und mitteln und erhält damit den Kreuzvalidierungsfehler CVE .

Overfitting vermeiden

Das Beispiel zeigt, dass sich die Prognosegüte – gemessen über den Kreuzvalidierungsfehler – verschlechtern kann, wenn zusätzliche Prädiktoren hinzugenommen werden, deren Aufklärungsbeitrag nicht signifikant ist. Dies kann man sich wie folgt erklären: Ist das Regressionsgewicht einer unabhängigen Variablen in der Population gleich 0, wird es in der Stichprobe aufgrund des Stichprobenfehlers aber typischerweise abweichend von 0 geschätzt. Dann führt die Gewichtung des Wertes der unabhängigen Variablen mit dem geschätzten – von 0 verschiedenen – Regressionsgewicht zu einer Verzerrung der Schätzung. Diese Verzerrung wirkt sich aufgrund des Stichprobenfehlers in unterschiedlichen Stichproben unterschiedlich aus. Daher sollte man es vermeiden, Prädiktoren in die Gleichung aufzunehmen, die in der Population ein Regressionsgewicht von 0 haben. Eine Missachtung dieser Empfehlung führt zu einem sog. »overfitting« (Überanpassung). Man hat mehr unabhängige Variablen im Modell, als benötigt werden, und dies bewirkt in aller Regel eine Verschlechterung der Prognosegüte.

19.7.8 Schätzung und Überprüfung des Achsenabschnitts β_0

Auch für den Populations-Achsenabschnitt β_0 ist der Achsenabschnitt b_0 , den man mittels der Kleinst-Quadrat-Schätzung anhand der Stichprobendaten nach Formel F 19.14 bestimmen kann, ein erwartungstreuer

Schätzer. Hypothesen über den Achsenabschnitt werden allerdings selten getestet. Der Standardfehler des Achsenabschnitts soll daher nicht im Detail beschrieben werden. Er wird von Fahrmeir et al. (2010) ausführlich behandelt und von Statistikprogrammen routinemäßig berechnet. Bezeichnet man den geschätzten Standardfehler des Achsenabschnitts mit $\hat{\sigma}_{B_0}$, so erhält man zur Überprüfung der Nullhypothese $H_0: \beta_0 = \beta_{0(0)}$ bzw. einer ihrer gerichteten Varianten folgenden Wert der Prüfgröße, die einer t -Verteilung mit $df = n - k - 1$ Freiheitsgraden folgt:

$$t = \frac{b_0 - \beta_{0(0)}}{\hat{\sigma}_{B_0}} \quad (\text{F 19.52})$$

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

$$b_0 \pm t_{(1-\frac{\alpha}{2}; n-k-1)} \cdot \hat{\sigma}_{b_0} \quad (\text{F 19.53})$$

Beispiel

Wohlbefinden und Stimmungsregulationskompetenz: Achsenabschnitt

In unserem Anwendungsbeispiel, in dem das Wohlbefinden auf die beiden Stimmungsregulationskompetenzen zurückgeführt wurde ($n = 237$), beträgt der geschätzte Achsenabschnitt $b_0 = 1,801$ mit einem Standardfehler von $\hat{\sigma}_{b_0} = 0,194$. Das bedeutet: Der geschätzte Wohlbefindenswert einer Person, die sowohl hinsichtlich ihrer Fähigkeit zur Aufrechterhaltung guter Stimmung als auch ihrer Fähigkeit zur Verbesserung schlechter Stimmung eine Ausprägung von 0 hat, beträgt 1,801. Nun soll geprüft werden, ob dieser Wert signifikant von 0 verschieden ist. Das zweiseitige 95%-Konfidenzintervall umfasst die Werte $[1,419; 2,183]$. Die Überprüfung der Nullhypothese $H_0: \beta_0 = 0$ führt zu einem t -Wert von $t = 9,291$, der sich nur aufgrund des Rundungsfehlers von $1,801/0,194$ unterscheidet. Der Wert ist größer als der kritische t -Wert von $t_{(0,975; 234)} = 1,970$, und die Nullhypothese muss auf einem Signifikanzniveau von $\alpha = 0,05$ bei einem zweiseitigen Test verworfen werden.

19.7.9 Schätzung der bedingten Erwartungswerte und individuell prognostizierter Werte

Wie bei der einfachen linearen Regression können auch bei der multiplen Regressionsanalyse die bedingten Er-

wartungswerte, die individuellen prognostizierten Werte sowie deren Konfidenzintervalle geschätzt werden. Das Grundprinzip ist dasselbe wie bei den Verfahren, die wir in den Abschnitten 17.9.6 und 17.9.7 beschrieben haben. Allerdings sind die Formeln für die Bestimmung der Konfidenzintervalle komplexer und setzen Kenntnisse der Matrixalgebra voraus. Wir verzichten daher auf ihre Darstellung. Die Konfidenzintervalle lassen sich von Statistikprogrammen wie z. B. R oder SPSS berechnen, die die unteren und oberen Grenzen der Konfidenzintervalle für den geschätzten Erwartungswert und den prognostizierten individuellen Wert direkt in die Datendatei schreiben. Dies ist hilfreich, da wir bereits in den Abschnitten 17.9.6 und 17.9.7 gesehen haben, dass die Breite des Konfidenzintervalls von der Ausprägung der unabhängigen Variablen abhängt. Dies ist auch bei der multiplen Regressionsanalyse so. Für jede Konfiguration von Werten auf den unabhängigen Variablen ergeben sich unterschiedliche Intervallgrenzen, die sich aufgrund der Mehrdimensionalität auch nicht mehr einfach grafisch darstellen lassen. Für die Praxis der Datenanalyse reicht es aus zu wissen, was die Grenzen der Konfidenzintervalle bedeuten und dass

sie für Individuen, die sich in ihren Werten auf den unabhängigen Variablen unterscheiden, verschieden sind.

19.8 Suppressorvariable

Wir haben in den vorherigen Abschnitten gesehen, dass sich der Einfluss einer unabhängigen Variablen auf die Kriteriumsvariable verringern kann, wenn weitere unabhängige Variablen in die Regressionsgleichung aufgenommen werden, die mit der ursprünglichen unabhängigen Variablen korreliert sind. Im Rahmen der multiplen Regressionsanalyse kann es jedoch nicht nur zu einer Verminderung des Einflusses einer unabhängigen Variablen auf die abhängige Variable kommen, wenn andere unabhängige Variablen in die Regressionsgleichung aufgenommen werden. Es kann auch der umgekehrte Fall eintreten, dass der Erklärungsbeitrag einer unabhängigen Variablen im Rahmen einer multiplen Regressionsanalyse größer ist als in einer einfachen Regressionsanalyse. Ein solches Beispiel haben wir schon im Kapitel zur Partialkorrelation kennengelernt. Wir werden es hier erneut aufgreifen.

Beispiel

Habituelle Stimmung, Momentanstimung, Stimmungsabweichung

Das Beispiel bezog sich auf die Vorhersage der habituellen Stimmung (*HS*) aus der momentanen guten (positiven) Stimmung (*PS*) und der Stimmungsabweichung (*SA*) (vgl. Tab. 18.1 c in Abschn. 18.1). In dem dort beschriebenen Datenbeispiel gab es zwischen der momentanen positiven Stimmung und der habituellen Stimmung eine Korrelation von $r = 0,307$, die anzeigt, dass man aufgrund der momentanen Stimmung mit einem gewissen Prognosefehler das habituelle Stimmungserleben einer Person vorhersagen kann. Außerdem haben wir gesehen, dass die momentane Stimmungsabweichung einer Person, d.h. ihre Angabe, ob sie sich in einer bestimmten Situation besser oder schlechter fühlt als im Allgemeinen, nicht mit der habituellen Stimmung korreliert ist. Die Korrelation zwischen der momentanen Stimmungsabweichung und dem habituellen Stimmungserleben betrug $r = -0,041$ und war nicht signifikant. Dies bedeutet, dass man aufgrund der momentanen Stimmungsabweichung nicht das habituelle Stimmungserleben einer Person prognostizieren kann, was im Einklang

mit theoretischen Überlegungen zur emotionalen Befindlichkeit steht. Nimmt man nun die momentane Stimmungsabweichung als einzigen Prädiktor in eine einfache lineare Regressionsanalyse auf, so wird das Regressionsgewicht dieser Variablen mit $b_1 = -0,036$ geschätzt und ist auf einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ nicht bedeutsam von 0 verschieden ($p = 0,592$). Die momentane Stimmungsabweichung leistet somit keinen Beitrag zur Prognose des habituellen Stimmungserlebens. Die Situation ändert sich jedoch, wenn man sowohl die momentane positive Stimmung als auch die Stimmungsabweichung als Prädiktoren zur Vorhersage der habituellen Stimmung betrachtet. Die Regressionsgleichung lautet wie folgt:

$$\widehat{HS} = 8,105 - 0,489 \cdot SA + 0,631 \cdot PS$$

Alle Regressionsgewichte sind auf einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ signifikant von 0 verschieden (alle $p < 0,01$). Zusammen erklären die beiden unabhängigen Variablen 24 % der Varianz des

habituellen Stimmungserlebens. Die positive Stimmung allein erklärt nur 9,4% der Varianz (wenn man sie als alleinige unabhängige Variable betrachtet). Über die positive Stimmung hinaus erklärt die Stimmungsabweichung zusätzlich 14,6% der Varianz. Ihr Semipartialdeterminationskoeffizient beträgt

$R^2_{(SA \cdot PS)HS} = 0,146$, obwohl ihr Determinationskoeffizient in der einfachen Regressionsanalyse nur $R^2 = 0,002$ beträgt. Durch die Hinzunahme der positiven Momentanstimung steigt der Erklärungsbeitrag der Stimmungsabweichung an. Dieser Zuwachs ist bedeutsam von 0 verschieden ($p < 0,01$).

Klassische Suppressorvariable

Dieses Beispiel zeigt sehr anschaulich, dass eine unabhängige Variable, die mit dem Kriterium unkorreliert ist, im Rahmen einer multiplen Regressionsanalyse durchaus einen bedeutsamen Beitrag zur Vorhersage einer Kriteriumsvariablen leisten kann. Dieses zunächst kontraintuitiv erscheinende Phänomen wurde von Horst (1941) Suppression genannt. Unter einer Suppressorvariablen X_2 versteht Horst (1941) eine Variable, die mit dem Kriterium unkorreliert ist, mit einer anderen unabhängigen Variablen jedoch eine bedeutsame Korrelation aufweist. Hierdurch kommt es zu dem Phänomen, dass die quadrierte multiple Korrelation größer ist als die quadrierte bivariate Korrelation zwischen der Kriteriumsvariablen Y und der anderen Prädiktorvariablen X_1 . Formal gesehen lässt sich dieses Phänomen anhand der folgenden drei Gleichungen beschreiben:

$$\rho_{YX_2} = 0 \quad (\text{F 19.54})$$

$$\rho_{X_1X_2} \neq 0 \quad (\text{F 19.55})$$

$$P^2_{Y|X_1, X_2} > \rho^2_{YX_1} \quad (\text{F 19.56})$$

Inhaltlich lässt sich das Phänomen wie folgt erklären: Die Korrelation zwischen der Variablen X_1 und der Kriteriumsvariablen zeigt an, dass die unabhängige Variable einen Teil enthält, der mit dem Kriterium zusammenhängt. Der Korrelation der beiden unabhängigen Variablen ist zu entnehmen, dass sie teilweise etwas Gemeinsames erfassen. Die Suppressorvariable X_2 enthält jedoch keinen Teil, der zur Vorhersage des Kriteriums geeignet ist, da ihre Korrelation mit dem Kriterium gleich 0 ist. Aufgrund ihrer Korrelation mit der Variablen X_1 ist die Variable X_2 in der Lage, einen Teil der Variablen X_1 zu »unterdrücken« (supprimieren), der nichts mit der Kriteriumsvariablen zu tun hat. Hierdurch wird der Vorhersagebeitrag der Variablen X_1 für die Kriteriumsvariable vergrößert.

In unserem Beispiel enthält die momentane Stimmung einen Anteil, der auf das habituelle Stimmungserleben zurückgeführt werden kann, und einen Teil, der spezifisch für eine Situation ist. Dieser spezifische Anteil, der auf die Situation zurückgeführt werden kann, ist für die Vorhersage des habituellen Wohlbefindens wenig brauchbar, da das habituelle Wohlbefinden ja gerade als situationsübergreifende Eigenschaft einer Person definiert ist. Aufgrund der momentanen Stimmungsabweichung lässt sich nun dieser situationsspezifische Anteil des momentanen Stimmungserlebens, der mit dem habituellen Wohlbefinden nichts zu tun hat, teilweise unterdrücken. Dadurch tritt derjenige Anteil der momentanen Stimmung klarer hervor, der indikativ für das habituelle Wohlbefinden ist. Das von Horst (1941) beschriebene Phänomen wird in der statistischen Literatur als *klassische Suppressorsituation* bezeichnet. Eine Variable X_2 , die die Anforderungen der Gleichungen F 19.54 bis F 19.56 erfüllt, heißt *klassische Suppressorvariable*. Hiervon lassen sich andere Suppressionskonzepte abgrenzen.

erleben zurückgeführt werden kann, und einen Teil, der spezifisch für eine Situation ist. Dieser spezifische Anteil, der auf die Situation zurückgeführt werden kann, ist für die Vorhersage des habituellen Wohlbefindens wenig brauchbar, da das habituelle Wohlbefinden ja gerade als situationsübergreifende Eigenschaft einer Person definiert ist. Aufgrund der momentanen Stimmungsabweichung lässt sich nun dieser situationsspezifische Anteil des momentanen Stimmungserlebens, der mit dem habituellen Wohlbefinden nichts zu tun hat, teilweise unterdrücken. Dadurch tritt derjenige Anteil der momentanen Stimmung klarer hervor, der indikativ für das habituelle Wohlbefinden ist. Das von Horst (1941) beschriebene Phänomen wird in der statistischen Literatur als *klassische Suppressorsituation* bezeichnet. Eine Variable X_2 , die die Anforderungen der Gleichungen F 19.54 bis F 19.56 erfüllt, heißt *klassische Suppressorvariable*. Hiervon lassen sich andere Suppressionskonzepte abgrenzen.

Andere Typen von Suppressorvariablen

Conger (1974) versteht unter einer Suppressorvariablen eine Variable X_2 , deren Aufnahme in die Regressionsanalyse dazu führt, dass das standardisierte multiple Regressionsgewicht einer anderen unabhängigen Variablen X_1 größer ist als die bivariate Korrelation (also das standardisierte bivariate Regressionsgewicht) zwischen der unabhängigen Variablen X_1 und der Kriteriumsvariablen Y . Formal lässt sich das durch folgende Ungleichung darstellen:

$$|\beta_{1s}| > |\rho_{YX_1}| \quad (\text{F 19.57})$$

Diese Definition verdeutlicht, dass durch die Suppressorvariable X_2 der Erklärungsbeitrag einer anderen Variablen X_1 vergrößert wird. Das Phänomen der klassischen Suppression lässt sich auch unter dieses Konzept der Definition einer Suppressorvariablen subsumieren. Tzelgov und Henik (1991) sowie Smith et al. (1992) diskutieren Bedingungen, unter denen sich ein solchermaßen definierter Suppressionseffekt einstellt.

Das geschätzte standardisierte Regressionsgewicht für die Momentanstimung in der multiplen Regressionsgleichung beträgt $b_{1s} = 0,708$, während die geschätzte bivariate Korrelation mit $r = 0,307$ kleiner ist.

Reziproke Suppression. Neben der klassischen Suppression lässt sich auch die sog. reziproke Suppression unter diese Definition einer Suppressorvariablen subsumieren. Unter reziproker Suppression versteht man das Phänomen, dass zwei unabhängige Variablen jeweils positiv mit der Kriteriumsvariablen korrelieren, jedoch untereinander eine negative Korrelation aufweisen (Conger, 1974). Dies führt dazu, dass das standardisierte multiple Regressionsgewicht der Variablen X_1 größer ist als ihre bivariate Korrelation mit dem Kriterium und

dass auch das standardisierte multiple Regressionsgewicht der Variablen X_2 größer ist als ihre bivariate Korrelation mit dem Kriterium. Das Konzept der reziproken Suppression lässt sich wie folgt formal definieren:

$$\rho_{YX_1} > 0, \quad \rho_{YX_2} > 0, \quad \rho_{X_1X_2} < 0 \quad (\text{F 19.58})$$

und

$$\beta_{1s} > \rho_{YX_1}, \quad \beta_{2s} > \rho_{YX_2} \quad (\text{F 19.59})$$

Reziproke Suppression bedeutet, dass die unabhängigen Variablen Anteile erfassen, die nichts mit der Kriteriumsvariablen zu tun haben. Dies erkennt man an der negativen Korrelation zwischen den unabhängigen Variablen. Gleichzeitig sind beide unabhängigen Variablen mit der Kriteriumsvariablen aber positiv korreliert.

Beispiel

Emotionale Selbstaufmerksamkeit

In einer Studie zur emotionalen Selbstaufmerksamkeit haben wir drei Skalen eingesetzt ($n = 240$):

- (1) eine Skala zur Erfassung der *allgemeinen emotionalen Selbstaufmerksamkeit* (ESA; Lischetzke et al., 2001), die die allgemeine Tendenz erfasst, auf eigene Gefühle zu achten;
- (2) die Skala *Dysfunktionale Selbstaufmerksamkeit* (DSA; Hoyer, 2000), die maladaptive Aspekte der Selbstaufmerksamkeit erfasst wie z.B. die Tendenz, nicht mehr aufhören zu können, auf den eigenen Zustand zu fokussieren, selbst wenn dies nicht zu einer Problemlösung beiträgt;
- (3) die Skala *Funktionale Selbstaufmerksamkeit* (FSA; Hoyer, 2000), die adaptive Aspekte der Selbstaufmerksamkeit erfasst, wie z.B. die Selbstwirksamkeitserwartung, eine Lösung für ein Problem zu finden und die Selbstaufmerksamkeit flexibel und zieladäquat steuern zu können.

Sowohl die dysfunktionale als auch die funktionale Selbstaufmerksamkeit sind positiv mit der allgemeinen emotionalen Selbstaufmerksamkeit korreliert

($r = 0,248$ bzw. $r = 0,192$), wohingegen die dysfunktionale und die funktionale Selbstaufmerksamkeit negativ zu $r = -0,343$ korreliert sind. Führt man nun eine multiple Regressionsanalyse mit der emotionalen Selbstaufmerksamkeit als abhängiger und sowohl der dysfunktionalen als auch der funktionalen Selbstaufmerksamkeit als unabhängigen Variablen durch, ergeben sich standardisierte Regressionsgewichte für die beiden unabhängigen Variablen, die größer sind als ihre bivariaten Korrelationen mit der abhängigen Variablen (s. Tab. 19.9). Außerdem ist der multiple Determinationskoeffizient größer als die Summe der Determinationskoeffizienten, die sich in den beiden einfachen Regressionsanalysen ergeben. Die Semipartialdetermination jeder der beiden unabhängigen Variablen mit der abhängigen Variablen ist größer als ihre quadrierte Korrelation mit dieser. Jeder der beiden Prädiktoren profitiert in Bezug auf seine Vorhersageleistung also von der Anwesenheit des anderen Prädiktors in der multiplen Regressionsanalyse.

Tabelle 19.9 Reziproke Suppression: Regression der emotionalen Selbstaufmerksamkeit (ESA) auf die dysfunktionale (DSA) und die funktionale (FSA) Selbstaufmerksamkeit

Unabhängige Variable	b_j	b_{js}	R	R^2
Einfache lineare Regressionsanalysen				
Dysfunktionale Selbstaufmerksamkeit	0,246	0,248	0,248	0,062
Funktionale Selbstaufmerksamkeit	0,213	0,192	0,192	0,037

Tabelle 19.9 (Fortsetzung)

Unabhängige Variable	b_j	b_{js}	R	R^2
Multiple Regressionsanalyse				
Dysfunktionale Selbstaufmerksamkeit	0,353	0,356	0,386	0,149
Funktionale Selbstaufmerksamkeit	0,349	0,314		
Korrelation				
$r_{DSA, FSA} = -0,343$				

Anmerkung: Alle angegebenen Größen sind auf einem $\alpha = 0,05$ signifikant.

Negative Suppression. Als dritte Form der Suppression lässt sich die negative Suppression definieren. Sind beide Variablen X_1 und X_2 so skaliert, dass sie positiv mit dem Kriterium korreliert sind, und ist darüber hinaus die Korrelation der beiden Variablen X_1 und X_2 ebenfalls positiv, dann ist die Variable X_2 eine Suppressorvariable, wenn ihre Korrelation mit der Kriteriumsvariablen kleiner ist als das Produkt aus der Korrelation

der Variablen X_1 mit dem Kriterium und der Korrelation der beiden unabhängigen Variablen untereinander. Formal ausgedrückt:

$$\rho_{X_1 X_2} > 0, \quad \rho_{Y X_2} < \rho_{Y X_1} \cdot \rho_{X_1 X_2} \quad (\text{F 19.60})$$

In diesem Fall bekommt die Suppressorvariable ein negatives Regressionsgewicht.

Beispiel

Vermögen, Schulden und finanzielle Zufriedenheit

Schätzt man die Populationskorrelationen mit den Stichprobenkorrelationen, so ist diese Situation im Beispiel in Tabelle 18.1 d (in Abschn. 18.1) erfüllt. Will man die finanzielle Zufriedenheit (Y) anhand des Vermögens (X_1) und der Schulden (X_2) vorher-sagen, so erhält man folgende Konstellation der Korrelationen:

$$\begin{aligned} r_{X_1 X_2} &> 0, & r_{Y X_2} &< r_{Y X_1} \cdot r_{X_1 X_2} \\ 0,346 &> 0, & 0,108 &< 0,165 \end{aligned}$$

In Tabelle 19.10 berichten wir die Ergebnisse der einfachen und der multiplen Regressionsanalyse. Diese zeigen, dass das Regressionsgewicht der Schulden sein Vorzeichen wechselt. Während in der einfachen Regressionsanalyse die Schulden einen positiven Einfluss auf die finanzielle Zufriedenheit haben, ist ihr Einfluss in der multiplen Regressionsanalyse negativ. Dies ist darauf zurückzuführen, dass in der einfachen Regressionsanalyse der Einfluss des Vermögens, das hoch positiv mit den Schulden korreliert ist, nicht kontrolliert wurde. Hohe Schulden zu haben scheint in der bivariaten Regression also mit größerer finanzieller

Zufriedenheit einherzugehen, aber die inhaltliche Bedeutung dieses Effekts wird erst durch die multiple Regressionsanalyse klar: Erst wenn man das Vermögen auspartialisiert, zeigt sich der wesentlich plausiblere negative Zusammenhang zwischen Schulden und Zufriedenheit: Hohe Schulden sind offenbar nur deshalb bivariat positiv mit der Zufriedenheit korreliert, weil Personen, die viele Schulden machen, generell mehr Geld besitzen. Es ist dieses Vermögen, das die Zufriedenheit ausmacht. Kontrolliert man jedoch den Einfluss des Vermögens (d.h., partialisiert man das Vermögen aus), so zeigt sich: Je höher die Schulden, desto geringer die finanzielle Zufriedenheit.

Die Variable Schulden ist darüber hinaus eine Suppressorvariable für die Variable Vermögen. Sie unterdrückt einen Anteil am Vermögen, der nichts mit der finanziellen Zufriedenheit zu tun hat, und erhöht dadurch die Vorhersageleistung des Vermögens: Das standardisierte Regressionsgewicht des Vermögens ist in der multiplen Regressionsanalyse größer als in der einfachen Regressionsanalyse.

Tabelle 19.10 Negative Suppression: Regression der finanziellen Zufriedenheit (FZ) auf das Vermögen (VE) und die Schulden (SC)

Unabhängige Variable	b_j	b_{js}	R	R^2
Einfache lineare Regressionsanalysen				
Vermögen	0,933	0,476	0,476	0,226
Schulden	$0,827 \cdot 10^{-7}$	0,108	0,108	0,012
Multiple Regressionsanalyse				
Vermögen	0,977	0,498	0,480	0,230
Schulden	$-0,5 \cdot 10^{-7}$	-0,065		

Anmerkung: Alle angegebenen Größen sind auf einem $\alpha = 0,05$ signifikant.

Suppression als Erhöhung der Nützlichkeit eines Prädiktors

Eine weitere Definition von Suppressorvariablen stammt von Velicer (1978). Dieser Definition zufolge liegt eine Suppressionssituation dann vor, wenn die Nützlichkeit einer Variablen X_j größer ist als die quadrierte Korrelation dieser Prädiktorvariablen mit dem Kriterium, formal ausgedrückt:

$$U_{X_j} > \rho_{YX_j}^2 \quad (\text{F } 19.61)$$

U_{X_j} bezeichnet die Nützlichkeit (engl. utility) der Variablen X_j , die wir in Abschnitt 19.6 definiert haben. Sie kennzeichnet den Anteil an zusätzlicher Varianzaufklärung, den die Variable X_j leistet, wenn man sie als letzte Variable in die Regressionsgleichung aufnimmt. Der Vorteil von Velicers Definition besteht darin, dass sie auch auf regressionsanalytische Modelle angewendet werden kann, in denen mehr als nur zwei unabhängige Variablen berücksichtigt werden. Einen Überblick über verschiedene Suppressionsdefinitionen und ihre Bedeutung für die psychologische Forschung geben Tzelgov und Henik (1991). Mit der Frage von Suppressionseffekten in der multiplen Regressionsanalyse und im Allgemeinen Linearen Modell setzen sich Holling (1983) sowie Smith et al. (1992) auseinander.

19.9 Moderierte Regressionsanalyse

In den bisher behandelten Regressionsanalysen wird angenommen, dass die unabhängigen Variablen additiv verknüpft sind. Wir haben in Abbildung 19.4 (in Abschn. 19.4.1) eine wichtige Implikation dieser additiven Verknüpfung gesehen: Das Regressionsgewicht einer unabhängigen Variablen hängt nicht von den Ausprägungen der anderen Variablen ab. Dies bedeutet, dass der Einfluss einer unabhängigen Variablen auf die Kriteriumsvariable für jede Ausprägung der anderen unabhängigen Variablen die gleiche Form aufweist. Es liegt somit keine Interaktion vor. Im Beispiel mit zwei unabhängigen Variablen in Abbildung 19.4 zeigt sich dies darin, dass die Regressionsgeraden parallel verlaufen. Diese rein additive Verknüpfung von unabhängigen Variablen ist in vielen Fällen zu einfach, wenn nicht gar psychologisch unplausibel. Welchen Effekt eine unabhängige Variable auf eine abhängige Variable hat, hängt nicht selten von der Ausprägung auf einer oder mehreren weiteren Variablen (sog. Moderatorvariablen) ab. Dies wollen wir am Beispiel der Abhängigkeit des habituellen Wohlbefindens von der Extraversion und von der Fähigkeit zur Ansteckung mit positiven Affekten erläutern.

Beispiel

Emotionale Ansteckung, Extraversion und Wohlbefinden

Unter der Fähigkeit zur Ansteckung mit positiven Affekten verstehen wir die Fähigkeit einer Person, positive Stimmungen aus ihrer Umgebung aufzunehmen

und für die Wohlbefindensregulation zu nutzen. Wir nennen sie im Folgenden einfach »emotionale Ansteckung«. Je stärker diese Fähigkeit ausgebildet ist,

umso stärker wird das habituelle Wohlbefinden einer Person ausgeprägt sein. In einer einfachen linearen Regressionsanalyse würden wir ein positives Regressionsgewicht erwarten. Welches Ergebnis würden wir erwarten, wenn wir zusätzlich die Extraversion als unabhängige Variable in das Modell aufnehmen? Man kann sich vorstellen, dass die Extraversion und die emotionale Ansteckung nicht rein additiv auf das Wohlbefinden wirken, sondern die Interaktion der Extraversion und der emotionalen Ansteckung darüber hinaus eine Rolle spielt. Interaktion bedeutet, dass der Einfluss der emotionalen Ansteckung von der Ausprägung der Extraversion abhängt bzw. umgekehrt dass der Einfluss der Extraversion von der Ausprägung der emotionalen Ansteckung abhängt. Warum? Man kann sich vorstellen, dass sich die emotionale Ansteckung bei Personen mit hoher Extraversion stärker auf das Wohlbefinden auswirkt als bei Personen mit niedriger Extraversion, da extravertierte Personen soziale Kontexte häufiger aufsuchen und daher häufiger in Situationen gelangen, die emotionale Ansteckung ermögli-

chen. Verfügt man über die Fähigkeit zur emotionalen Ansteckung, aber meidet tendenziell soziale Kontakte (was bei eher introvertierten Personen der Fall ist), so wird sich diese auf das Wohlbefinden nicht auswirken können. In der Gruppe der hoch Extravertierten würde man daher eine steilere Regressionsgerade, die den Zusammenhang zwischen dem Wohlbefinden und der emotionalen Ansteckung beschreibt, erwarten als in einer Gruppe von Introvertierten. Da es sich bei der Extraversion nicht um eine dichotome, sondern im Allgemeinen um eine metrische Variable handelt, kann man diese Hypothese auch stärker formulieren, indem man annimmt, dass das Regressionsgewicht, das den Zusammenhang zwischen Wohlbefinden und der emotionalen Ansteckung beschreibt, eine Funktion der Extraversion ist. Im Folgenden werden wir zeigen, wie solche Moderatorhypothesen im Rahmen der multiplen Regressionsanalyse überprüft werden können, wobei wir uns auf das Beispiel der Abhängigkeit des Wohlbefindens von der Extraversion und der emotionale Ansteckung beziehen werden.

Eine Regressionsanalyse, in der der Einfluss einer unabhängigen Variablen von der Ausprägung einer anderen unabhängigen Variablen abhängt, heißt moderierte Regressionsanalyse. In Abschnitt 4.1.3 haben wir das Konzept einer Moderatorvariablen schon vorgestellt, das sich auf die Regressionsanalyse übertragen lässt.

19.9.1 Moderierte Regressionsanalyse: Zwei unabhängige Variablen

Das Grundmodell der moderierten Regressionsanalyse für eine abhängige Variable Y und zwei unabhängige Variablen X_1 und X_2 lässt sich auf Ebene der Stichprobe wie folgt formulieren:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_1 \cdot X_2 + E \quad (\text{F 19.62})$$

Das Modell der moderierten Regression unterscheidet sich von dem bisherigen Modell der multiplen Regressionsanalyse dahin gehend, dass eine dritte unabhängige Variable in die Regressionsgleichung aufgenommen wird, das Produkt der beiden unabhängigen Variablen X_1 und X_2 . Die Werte dieser Produktvariablen lassen sich einfach berechnen, indem man für jede Person ihre Werte auf den beiden unabhängigen Variablen multi-

pliziert. Diese Produktvariable wird dann wie eine ganz gewöhnliche unabhängige Variable zusätzlich in die Regressionsgleichung mit aufgenommen. Einige Statistikprogramme wie z. B. SPSS übernehmen die Berechnung der Produktvariablen automatisch, wenn man im Rahmen der Option zum Allgemeinen Linearen Modell ein Modell mit einer Interaktionsvariablen spezifiziert. Die vorhergesagten Werte \hat{y}_m hängen somit von den Werten auf den beiden unabhängigen Variablen sowie ihrem Produkt ab:

$$\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_1 \cdot X_2 \quad (\text{F 19.63})$$

Will man wie in Abbildung 19.4 die regressive Abhängigkeit der abhängigen Variablen Y von einer der beiden unabhängigen Variablen (z. B. X_1) – bei Konstanzhaltung der anderen unabhängigen Variablen (z. B. X_2) – darstellen, muss man Gleichung F 19.63 wie folgt umformen:

$$\begin{aligned} \hat{Y} &= \underbrace{(b_0 + b_2 \cdot X_2)}_{b_{01}} + \underbrace{(b_1 + b_3 \cdot X_2)}_{b_{11}} \cdot X_1 \\ \hat{Y} &= b_{01} + b_{11} \cdot X_1 \end{aligned} \quad (\text{F 19.64})$$

Diese Darstellungsform der moderierten Regression zeigt, dass die abhängige Variable Y in linearer Weise

von der unabhängigen Variablen X_1 abhängt, dass die Regressionskoeffizienten b_{01} und b_{11} jedoch ihrerseits lineare Funktionen der anderen unabhängigen Variablen darstellen: Je nachdem, welchen Wert die Variable X_2 annimmt, ändern sich auch die Regressionskoeffizienten b_{01} und b_{11} . Für einen bestimmten Wert x_2 der Variablen X_2 lässt sich die Abhängigkeit der Variablen Y von X_1 darstellen als:

$$\hat{Y} = (b_0 + b_2 \cdot x_2) + (b_1 + b_3 \cdot x_2) \cdot X_1 \quad (\text{F 19.65})$$

Bei dieser Darstellung wird deutlicher, dass es für einen festgelegten x_2 -Wert einen festen Achsenabschnitt und ein festes Regressionsgewicht gibt. Ist $x_2 = 0$, so erhält man die Gleichung:

$$\hat{Y} = b_0 + b_1 \cdot X_1$$

Das Regressionsgewicht b_1 in der moderierten Regressionsanalyse entspricht somit der Steigung der Regressionsgeraden für die Regression von Y auf X_1 für alle Personen, die auf der Variablen X_2 einen Wert von 0 aufweisen.

Man könnte die Regressionsgleichung natürlich auch so umstellen, dass die Abhängigkeit der Variablen Y von X_2 – bei gegebenem Wert von X_1 – betrachtet wird. Dann erhält man:

$$\hat{Y} = (b_0 + b_1 \cdot x_1) + (b_2 + b_3 \cdot x_1) \cdot X_2$$

In analoger Weise entspricht das Regressionsgewicht b_2 in der moderierten Regressionsanalyse der Steigung der Regressionsgeraden für die Regression von Y auf X_2 für alle Personen, die auf der Variablen X_1 einen Wert von 0 aufweisen.

Beispiel

Emotionale Ansteckung, Extraversion und Wohlbefinden: Moderierte Regression

Wir wollen die moderierte Regressionsanalyse anhand einer empirischen Anwendung illustrieren. Anhand einer Stichprobe von $n = 137$ Studierenden wurden das habituelle Wohlbefinden (Y : WB), die Extraversion (X_1 : EX) und die emotionale Ansteckung (X_2 : EA) erfasst. Das Modell der moderierten Regressionsanalyse lässt sich wie folgt formulieren:

$$\widehat{WB} = b_0 + b_1 \cdot EX + b_2 \cdot EA + b_3 \cdot EX \cdot EA \quad (\text{F 19.66})$$

Die Extraversion- und die emotionalen Ansteckungswerte wurden für alle Personen multipliziert, und die so gewonnene Produktvariable wurde als dritte unabhängige Variable in die Regressionsanalyse aufgenommen. Die Analyse ergab die folgenden geschätzten Regressionskoeffizienten:

$$\widehat{WB} = 4,380 - 0,276 \cdot EX - 0,382 \cdot EA + 0,147 \cdot EX \cdot EA$$

Zur Illustration betrachten wir die Extraversion als Moderatorvariable und stellen die Regression des Wohlbefindens auf die emotionale Ansteckung in Ab-

hängigkeit von den Extraversionswerten dar. Durch Umformung der Gleichung nach der allgemeinen Gleichung F 19.65 erhalten wir:

$$\begin{aligned} \widehat{WB} &= (4,380 - 0,276 \cdot EX) \\ &\quad + (-0,382 + 0,147 \cdot EX) \cdot EA \end{aligned}$$

Man sieht an dieser Gleichung, dass sowohl der Achsenabschnitt als auch der Steigungskoeffizient eine Funktion der Extraversion sind. Je größer der Extraversionswert ist, desto kleiner ist der Achsenabschnitt und desto größer ist der Anstieg der Regressionsgeraden. Grafisch stellt man die moderierte Regression häufig derart dar, dass die regressive Beziehung für drei ausgewählte Werte der Moderatorvariablen gezeichnet wird. Üblicherweise wählt man hierzu (1) den Mittelwert der Moderatorvariablen aus, (2) den Wert, der eine Standardabweichungseinheit unter dem Mittelwert liegt, und (3) den Wert, der eine Standardabweichungseinheit über dem Mittelwert liegt.

In Abbildung 19.6 sind die Regressionsgeraden für die drei entsprechenden Ausprägungen der Extraversion dargestellt, und zwar erstens für den Mittelwert der Extraversion (3,789), zweitens für den Wert, der eine Standardabweichungseinheit unter dem Mittelwert der Extraversion liegt ($3,789 - 0,772 = 3,017$) und drittens für den Wert, der eine Standardabweichungseinheit über dem Mittelwert der Extraversion liegt ($3,789 + 0,772 = 4,561$). Die grafische Veranschaulichung der drei Geraden zeigt, dass der Steigungskoeffizient mit Zunahme der Extraversion größer wird, die Fähigkeit zur Ansteckung mit positiver Stimmung sich somit umso stärker auf das Wohlbefinden auswirkt, je größer die Extraversion einer Person ist. In dieser Grafik sind nur drei Regressionsgeraden dargestellt; es ist jedoch wichtig darauf hinzuweisen, dass es eine solche Regressionsgerade für jeden Wert der Moderatorvariablen gibt, also für jeden Extraversionswert.

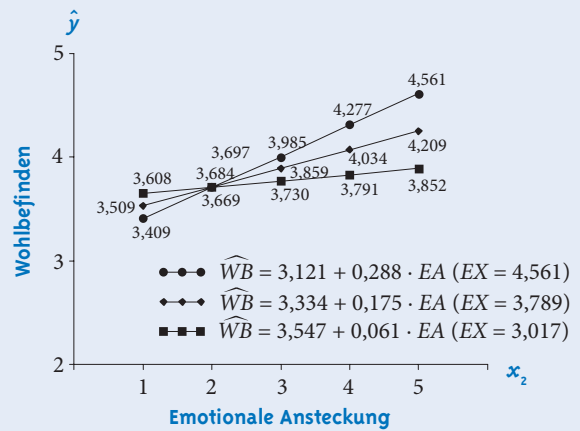


Abbildung 19.6 Darstellung der moderierten Regression: Regression des Wohlbefindens auf die emotionale Ansteckung für drei Extraversionswerte

19.9.2 Moderierte Regression mit zentrierten Variablen

Wir haben im letzten Abschnitt gesehen, dass sich die Regressionsgewichte b_1 und b_2 der Variablen X_1 bzw. X_2 in der moderierten Regressionsanalyse als Regressionsgewichte von einfachen Regressionen interpretieren lassen, bei denen die Ausprägung der anderen Variablen gleich 0 ist. Diese Interpretation ist in einer Anwendung jedoch nur dann bedeutungsvoll, wenn die Variablen jeweils den Wert 0 aufweisen können. Dies ist nicht in jeder Anwendung der Fall. In unserem Anwendungsbeispiel wurden die Skalen so gebildet, dass der minimal mögliche Wert gleich 1 ist, die 0 kann somit gar nicht vorkommen. Um die Interpretation der Regressionsgewichte zu erleichtern, können die unabhängigen Variablen zentriert werden, bevor man sie in die Regressionsanalyse aufnimmt. Unter Zentrierung versteht man, dass von jedem Messwert der Mittelwert der Variablen abgezogen wird. Für das Beispiel von zwei unabhängigen Variablen erhält man zentrierte Werte, indem man von jedem Wert x_{1m} den Mittelwert \bar{x}_1 abzieht und von jedem Wert x_{2m} den Mittelwert \bar{x}_2 abzieht. Die resultierenden Werte sind also nun um ihren ursprünglichen Mittelwert zentriert. Solche *zentrierten Variablen* werden auch als Abweichungsvariablen bezeichnet. Um zentrierte von unzentrierten Variablen zu unterscheiden, schreiben wir im Index einer zent-

rierten Variablen (oder einer Abweichungsvariablen) ein A (für Abweichung). Damit sind die beiden Abweichungsvariablen wie folgt definiert:

$$X_{1A} = X_1 - \bar{X}_1 \quad \text{und} \quad X_{2A} = X_2 - \bar{X}_2$$

Der Produktterm wird dann aus den zentrierten Variablen gebildet:

$$X_{1A} \cdot X_{2A} = (X_1 - \bar{X}_1) \cdot (X_2 - \bar{X}_2)$$

Wichtig ist, dass es sich hier um eine Produktvariable aus zentrierten Variablen handelt (und nicht etwa um ein zentriertes Produkt zweier unzentrierter Variablen!). Auch werden nur die unabhängigen Variablen und nicht die abhängige Variable zentriert.

Das vollständige Regressionsmodell im Falle von zentrierten Variablen lautet wie folgt:

$$\hat{Y} = b_0 + b_1 \cdot (X_1 - \bar{X}_1) + b_2 \cdot (X_2 - \bar{X}_2) + b_3 \cdot (X_1 - \bar{X}_1) \cdot (X_2 - \bar{X}_2) \quad (\text{F 19.67})$$

bzw.

$$\hat{Y} = b_0 + b_1 \cdot X_{1A} + b_2 \cdot X_{2A} + b_3 \cdot X_{1A} \cdot X_{2A} \quad (\text{F 19.68})$$

Im Falle zentrierter Variablen entspricht das Regressionsgewicht b_1 der Steigung der Regressionsgeraden von X_{1A} für eine mittlere Ausprägung der Variablen X_{2A} . Es lässt sich sogar zeigen, dass das Regressionsgewicht b_1 bei zentrierten Variablen die durchschnittliche Regres-

Tabelle 19.11 Korrelation zwischen einem Produktterm $X_1 \cdot X_2$ und seinen Konstituenten X_1 bzw. X_2 in Abhängigkeit von Transformationen von X_1 bzw. X_2

X_1	X_2	$X_1 \cdot X_2$		X_1	X_2
1	1	1			
2	3	6		X_2	0,48
3	5	15		$X_1 \cdot X_2$	0,81
4	1	4			0,84
5	3	15			
6	5	30			

X_1	X_2	$X_1 \cdot X_2$		X_1	X_2
21	1	21			
22	3	66		X_2	0,48
23	5	115		$X_1 \cdot X_2$	0,57
24	1	24			0,99
25	3	75			
26	5	130			

X_{1A}	X_{2A}	$X_{1A} \cdot X_{2A}$		X_{1A}	X_{2A}
-2,5	-2	5			
-1,5	0	0		X_{2A}	0,48
-0,5	2	-1		$X_{1A} \cdot X_{2A}$	0
0,5	-2	-1			0
1,5	0	0			
2,5	2	5			

sion über alle Ausprägungen der anderen unabhängigen Variablen hinweg beschreibt (Cohen et al., 2003). Dies gilt in analoger Weise für das Regressionsgewicht b_2 .

Die Zentrierung dient dazu, die Interpretation der Regressionsgewichte zu vereinfachen. Sie hat aber noch eine weitere Konsequenz. Sie verändert die Korrelation der Produktvariablen $X_1 \cdot X_2$ mit ihren Konstituenten X_1 bzw. X_2 .

Wie hoch die Produktvariable $X_1 \cdot X_2$ mit ihren Konstituenten X_1 bzw. X_2 korreliert ist, hängt zum einen von der Interkorrelation zwischen X_1 und X_2 ab: Je höher die

Korrelation der beiden unabhängigen Variablen, desto größer ist auch der Produktterm mit beiden von ihnen korreliert. Interessanterweise ändert sich die Korrelation der Produktvariablen mit ihren beiden Konstituenten, wenn man die Konstituenten in zulässiger Weise transformiert: So lässt sich zeigen, dass sich die Korrelation zwischen $X_1 \cdot X_2$ und X_2 erhöhen kann, wenn man eine Konstante zu X_1 hinzuaddiert, was bei metrischen Variablen zulässig ist (s. Abschn. 5.5). Das Gleiche gilt umgekehrt: Die Korrelation zwischen $X_1 \cdot X_2$ und X_1 kann sich erhöhen, wenn man eine Konstante zu X_2 hinzuaddiert.

In Tabelle 19.11 ist ein Beispiel dargestellt: Die beiden Variablen X_1 und X_2 im oberen Teil der Tabelle korrelieren miteinander zu $r = 0,48$. Der Produktterm $X_1 \cdot X_2$ korreliert mit X_1 zu $r = 0,81$ und mit X_2 zu $r = 0,84$. Addiert man nun eine Konstante von 20 auf X_1 (s. den mittleren Teil von Tab. 19.11), so erhöht sich natürlich nicht die Korrelation zwischen X_1 und X_2 (denn der Korrelationskoeffizient von Pearson ist invariant gegenüber linearen Transformationen), wohl aber die Korrelation zwischen $X_1 \cdot X_2$ und X_2 : Sie beträgt als Folge der Transformation nun $r = 0,99$.

Zentriert man die Werte, erhält man die entsprechenden Ergebnisse im unteren Teil von Tabelle 19.11. Es lässt sich leicht feststellen, dass nun sowohl X_{1A} als auch X_{2A} jeweils einen Mittelwert von 0 haben. Allein die Zentrierung hat dazu geführt, dass die Korrelation zwischen $X_{1A} \cdot X_{2A}$ und X_{1A} sowie zwischen $X_{1A} \cdot X_{2A}$ und X_{2A} restlos verschwindet: Sie ist gleich 0. Die Korrelation zwischen X_{1A} und X_{2A} beträgt weiterhin $r = 0,48$. Die Korrelation zwischen einem Produktterm $X_{1A} \cdot X_{2A}$ und seinen zentrierten Konstituenten X_{1A} und X_{2A} ist immer dann 0, wenn beide unabhängigen Variablen normalverteilt sind. Je asymmetrischer X_{1A} und X_{2A} verteilt sind, desto stärker weicht die Korrelation zwischen dem Produktterm und seinen zentrierten Konstituenten von 0 ab. Aber auch in diesem Fall ist die Korrelation deutlich geringer als im Falle unzentrierter Variablen (und ihres Produktterms).

Weist eine unabhängige Variable eine hohe multiple Korrelation mit den anderen unabhängigen Variablen auf, so liegt Multikollinearität vor (s. ausführlicher Abschn. 19.13.4). Während die Korrelation der unabhängigen Variablen (der Konstituenten des Produktterms)

von linearen Transformationen unberührt bleibt, ist die Korrelation des Produktterms mit den Konstituenten von der Skalierung abhängig. Wie in Tabelle 19.11 zu sehen ist, kann diese sehr hoch, aber auch 0 sein. Der Anteil der Multikollinearität der Produktvariablen, der allein von der Skalierung (z. B. lineare Transformation) abhängt, wird nicht-essenzielle Multikollinearität genannt. Die Multikollinearität, die nicht von der Skalierung der Variablen abhängt, wird essenzielle Multikollinearität genannt. Durch die Zentrierung kann die nicht-essenzielle Multikollinearität beeinflusst werden. Die Zentrierung ist aber nicht in der Lage, die essenzielle Multikollinearität zu beheben, die beispielsweise darauf zurückzuführen ist, dass die unabhängigen Variablen nicht normalverteilt sind. Während man in früheren Arbeiten davon ausgegangen ist, dass sich die Zentrierung grundsätzlich positiv auf die Verringerung der Multikollinearität auswirkt, hat z. B. Shieh (2011) gezeigt, dass eine Zentrierung auch zu einer Erhöhung der Multikollinearität führen kann. Da die inzwischen verfügbaren Computerprogramme mit hoher Präzision arbeiten, ist es eher unwahrscheinlich, dass sich durch die Verringerung der Multikollinearität aufgrund der Zentrierung rechentechnische Vorteile ergeben (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Shieh, 2011). Der Vorteil der Zentrierung ist daher vor allem in der besseren Interpretierbarkeit der Regressionsgewichte zu sehen. Die Zentrierung ist daher im Allgemeinen zu empfehlen, wenn der Wert 0 der Variablen keine sinnvolle Bedeutung hat. Liegen jedoch unabhängige Variablen vor, bei denen der Wert 0 eine sinnvolle Bedeutung hat, ist eine Zentrierung nicht zwingend notwendig.

Beispiel

Emotionale Ansteckung, Extraversion und Wohlbefinden: Zentrierte Variablen

In Tabelle 19.12 sind die geschätzten Parameter für die moderierte Regression auf der Grundlage der unzentrierten Variablen und der zentrierten Variablen zusammengestellt. Darüber hinaus werden auch die Korrelationen der unabhängigen Variablen einander gegenübergestellt. Diese Tabelle verdeutlicht einige wichtige Eigenschaften:

(1) Das Regressionsgewicht der Produktvariablen sowie der Standardfehler ihres Regressionsgewichts unterscheiden sich nicht zwischen den Modellen mit unzentrierten und zentrierten Variablen, d. h.,

die Größe des geschätzten Interaktionseffekts sowie seine inferenzstatistische Absicherung hängen nicht davon ab, ob die unabhängigen Variablen zentriert wurden oder nicht.

(2) Die Regressionsgewichte der unabhängigen Variablen und der Achsenabschnitt hängen davon ab, ob zentrierte oder unzentrierte Variablen gewählt werden.

(3) Die Standardfehler der geschätzten Regressionsparameter der unabhängigen Variablen und des Achsenabschnitts sind bei zentrierten Variablen

deutlich kleiner als bei unzentrierten Variablen. Dies ist darauf zurückzuführen, dass die unabhängigen Variablen mit der Produktvariablen bei zentrierten Variablen deutlich geringer korreliert sind als bei unzentrierten Variablen.

(4) Während die Korrelation der Produktvariablen bei unzentrierten Variablen mit der Extraversion $r = 0,804$ und mit der emotionalen Ansteckung $r = 0,828$ beträgt, sind die Korrelationen bei zentrierten Variablen mit $r = -0,170$ und $r = -0,142$ deutlich geringer.

Tabelle 19.12 Vergleich der geschätzten Regressionsparameter und deren Standardfehler einer moderierten Regression mit der abhängigen Variablen Wohlbefinden (WB) und den unabhängigen Variablen Extraversion (EX) und emotionale Ansteckung (EA) für zentrierte und unzentrierte Variablen

	Unzentrierte Variablen			Zentrierte Variablen		
	Regressionsparameter (Standardfehler)	t-Wert	p	Regressionsparameter (Standardfehler)	t-Wert	p
Konstante (Achsenabschnitt)	4,380 (0,805)	5,442	< 0,001	3,859 (0,040)	96,817	< 0,001
Extraversion (EX)	-0,276 (0,218)	-1,264	0,207	0,163 (0,053)	3,086	0,002
Emotionale Ansteckung (EA)	-0,382 (0,280)	-1,366	0,173	0,176 (0,064)	2,776	0,006
Produktvariable (EX · EA)	0,147 (0,073)	2,007	0,046	0,147 (0,073)	2,007	0,046

Korrelationen

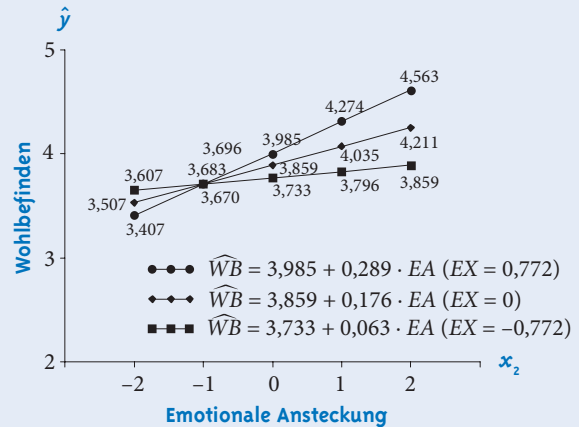
	Unzentrierte Variablen		Zentrierte Variablen	
	EX	EA	EX	EA
Extraversion (EX)				
Emotionale Ansteckung (EA)	0,356		0,356	
Produktvariable (EX · EA)	0,804	0,828	-0,170	-0,142

Wie man Abbildung 19.7 entnehmen kann, ist der Verlauf der bedingten Regressionsgeraden für die drei Werte der Extraversion unabhängig davon, ob die Regressionsanalyse mit zentrierten oder mit unzentrierten Variablen durchgeführt wurde. Die Regressionsgewichte unterscheiden sich nur aufgrund des Rundungsfehlers. Auch die Determinationskoeffizienten unterscheiden sich nicht zwischen beiden Ansätzen. Der Determinationskoeffizient beträgt in beiden Fällen $R^2 = 0,104$. In beiden Ansätzen ist auch der Zuwachs des Determinationskoeffizienten R^2 , der durch die Hinzunahme der Produktvariablen erreicht wird, mit 1,5 % gleich. Dieser zusätzliche Erklärungsgewinn mag auf den ersten Blick klein erscheinen. Es

xist hierbei jedoch zu beachten, dass die Erklärungskraft von Interaktionsvariablen im Allgemeinen recht klein ist. Nach Champoux und Peters (1987) liegt sie für psychologische Untersuchungen im Bereich von 3 %, nach Chaplin (1991) liegt sie im Bereich der differenzialpsychologischen Forschung bei etwa 8 %. Für die Vorhersage eines individuellen Kriteriumswertes ist der Moderatoreffekt jedoch von großer Bedeutung. So zeigt Abbildung 19.7, dass eine Person mit einem zentrierten Ansteckungswert von 2, die eine um eine Standardabweichung unterdurchschnittliche Extraversion aufweist, einen erwarteten Wohlbefindenswert von 3,86 zugeordnet bekommt, während dieser Wert für eine Person mit gleicher emotionaler

Ansteckung, aber mit einem um eine Standardabweichung überdurchschnittlichen Extraversionswert 4,58 beträgt. Dies ist für eine Variable mit einem Wertebereich von 1 bis 5 ein erheblicher Unterschied.

Abbildung 19.7 Darstellung der moderierten Regression: Regression des Wohlbefindens auf die emotionale Ansteckung für drei Extraversionswerte: Zentrierte Variablen



19.9.3 Inferenzstatistische Absicherung eines Moderatoreffekts

Ob in einer empirischen Anwendung ein Moderatoreffekt vorliegt oder nicht, lässt sich dadurch überprüfen, ob das Regressionsgewicht der Produktvariablen bedeutsam von 0 abweicht. In unserem Beispiel mit zwei unabhängigen Variablen lautet das Populationsmodell der moderierten Regression:

$$E(Y | X_1, X_2) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 \cdot X_2 \quad (\text{F 19.69})$$

Folglich lautet die Nullhypothese:

$$H_0: \beta_3 = 0$$

Die Alternativhypothese lautet:

$$H_1: \beta_3 \neq 0$$

Diese Hypothese kann auch gerichtet formuliert werden. Bei der inferenzstatistischen Absicherung des Interaktionseffekts ist zu beachten, dass dieser nur dann korrekt überprüft wird, wenn in der Regressionsgleichung nicht nur die Produktvariable, sondern auch die beiden unabhängigen Variablen, auf deren Grundlage die Produktvariable gebildet wurde, in der Regressionsgleichung enthalten sind. Dies ist notwendig, da die Produktvariable mit den beiden anderen unabhängigen Variablen korreliert sein kann, selbst wenn die Variablen zentriert sind. Würde man die unabhängigen Variablen nicht in die Regressionsgleichung aufnehmen, könnten sich hinter einem signifikanten Interaktionseffekt Effekte der beiden unabhängigen Variablen verbergen. Zur Überprüfung der Moderatorhypothese ist es daher notwendig zu zei-

gen, dass die Interaktionsvariable über die unabhängigen Variablen hinaus einen bedeutsamen Beitrag zur Erklärung der Variation in der Kriteriumsvariablen leistet.

Inferenzstatistisch kann die Interaktionshypothese auf zwei Arten überprüft werden: Zum einen lässt sich die Nullhypothese $H_0: \beta_3 = 0$ anhand eines t -Tests statistisch überprüfen. Die Ergebnisse in Tabelle 19.12 zeigen, dass in unserem Beispiel die Nullhypothese verworfen werden muss. Die zweite Möglichkeit besteht darin, zwei Regressionsmodelle miteinander zu vergleichen, wobei das erste (eingeschränkte) Regressionsmodell nur die beiden unabhängigen Variablen (z. B. Extraversion und Ansteckung mit positivem Affekt) enthält, das zweite (uneingeschränkte) Regressionsmodell daneben noch die Produktvariable. Inferenzstatistisch wird dann überprüft, ob die Hinzunahme der Interaktionsvariablen zu einer signifikanten Erhöhung der erklärten Varianz führt. Das Ergebnis dieses Tests ist in Tabelle 19.13 dargestellt. Er führt exakt zu demselben Ergebnis wie der Test des Regressionsgewichts anhand des t -Tests. Der F -Wert entspricht genau dem quadrierten t -Wert.

Bedingte Regressionsgewichte

Liegt ein bedeutsamer Moderatoreffekt vor und wählt man die Variable X_1 als Moderatorvariable aus, so unterscheiden sich die Regressionsgewichte von X_2 für verschiedene Ausprägungen von X_1 (s. Abb. 19.6 und 19.7). Bezogen auf das Populationsmodell sieht die bedingte lineare Regression wie folgt aus:

$$E(Y | X_1 = x_1, X_2) = (\beta_0 + \beta_1 \cdot x_1) + (\beta_2 + \beta_3 \cdot x_1) \cdot X_2$$

Ist man daran interessiert, die Nullhypothese, dass das bedingte lineare Regressionsgewicht $(\beta_2 + \beta_3 \cdot x_1)$

Tabelle 19.13 Determinationskoeffizienten und Veränderungen der Determinationskoeffizienten in dem Modell ohne (I) und mit Produktvariable (II)

Unabhängige Variable	R^2	Änderung in R^2	F-Wert	df_1	df_2	Signifikanz
Modell I						
Extraversion						
Ansteckung mit positivem Affekt	0,089					
Modell II						
Extraversion						
Ansteckung mit positivem Affekt						
Interaktion	0,104	0,015	4,028	1	233	0,046

gleich 0 ist, zu überprüfen, muss man auf spezifische Signifikanztests zurückgreifen. Im Folgenden bezeichnen wir dieses bedingte lineare Regressionsgewicht, das im Englischen auch *simple slope* genannt wird, mit β_{x_1} : $\beta_{x_1} = (\beta_2 + \beta_3 \cdot x_1)$. Es wird über das Regressionsgewicht $b_{x_1} = (b_2 + b_3 \cdot x_1)$ geschätzt. Sein Standardfehler lässt sich wie folgt bestimmen (Cohen et al., 2003; Preacher et al., 2006):

$$\hat{\sigma}_{B_{x_1}} = \sqrt{\hat{\sigma}_{B_2}^2 + 2 \cdot x_1 \cdot \hat{\sigma}_{B_2 B_3} + x_1^2 \cdot \hat{\sigma}_{B_3}^2} \quad (\text{F } 19.70)$$

In dieser Formel bezeichnen $\hat{\sigma}_{B_2}^2$ und $\hat{\sigma}_{B_3}^2$ die quadrierten Standardfehler der Regressionsgewichte und $\hat{\sigma}_{B_2 B_3}$ die Kovarianz der Schätzer der beiden Regressionsgewichte. Zur statistischen Überprüfung der Nullhypothese, dass ein bedingtes Regressionsgewicht (simple slope) an einer Stelle x_1 gleich 0 ist, teilt man das geschätzte Regressionsgewicht durch den Standardfehler und erhält eine t -verteilte Prüfgröße mit $df = n - k - 1$ Freiheitsgraden:

$$t = \frac{b_{x_1}}{\hat{\sigma}_{B_{x_1}}} \quad (\text{F } 19.71)$$

Konfidenzintervall. Ein zweiseitiges $(1 - \alpha)$ -Konfidenzintervall für das bedingte Regressionsgewicht ergibt sich dann wie folgt:

$$b_{x_1} \pm \hat{\sigma}_{B_{x_1}} \cdot t_{(1-\frac{\alpha}{2}; df)} \quad (\text{F } 19.72)$$

Für unser Beispiel in Abbildung 19.7 sind die entsprechenden Tests und geschätzten Konfidenzintervalle in Tabelle 19.14 zusammengestellt. Die Ergebnisse zeigen, dass die bedingten Regressionsgewichte für zwei Fälle von 0 verschieden sind, und zwar für einen mittleren Extraversionswert und einen Extraversionswert, der um eine Standardabweichung über dem Mittelwert liegt. Liegt der Extraversionswert jedoch eine Standardabweichung unter dem Mittelwert, so ist das bedingte Regressionsgewicht nicht signifikant von 0 verschieden. In dieser Gruppe gibt es somit keinen signifikanten Zusammenhang zwischen emotionaler Ansteckung und dem Wohlbefinden. Entsprechende Tests und Konfidenzintervalle lassen sich auch für $(\beta_0 + \beta_1 \cdot x_1)$ bestimmen, sind meist aber nicht von wissenschaftlichem Interesse. Preacher et al. (2006) haben ein im Internet frei verfügbares Tool entwickelt, das es erlaubt, die Teststatistiken und Konfidenzintervalle zu bestimmen.

Tabelle 19.14 Inferenzstatistische Tests und Konfidenzintervalle für die drei bedingten Regressionsgewichte in Abbildung 19.7

Ausprägung der Extraversion	b_{x_1}	$\hat{\sigma}_{B_{x_1}}$	t-Wert	p-Wert	95 %-Konfidenzintervall	
					Untergrenze	Obergrenze
EX = 0,772	0,289	0,076	3,802	< 0,01	0,139	0,439
EX = 0	0,176	0,064	2,770	< 0,01	0,050	0,302
EX = -0,772	0,063	0,093	0,670	0,504	-0,120	0,246

Luhmann (2013) zeigt, wie dies mit spezifischen Paketen des Computerprogramms R möglich ist.

Johnson-Neyman-Intervall. Die inferenzstatistische Überprüfung einzelner bedingter Regressionsgewichte hat den Nachteil, dass man die Ausprägungen der Moderatorvariablen vorher festlegen muss und den Test für diese spezifischen Werte durchführen muss. Häufig ist man jedoch weniger an einzelnen Werten der Moderatorvariablen interessiert, sondern an Bereichen der Moderatorvariablen, für welche die bedingten Regressionsgewichte signifikant von 0 verschieden sind. Hierzu kann man ein Johnson-Neyman-Intervall berechnen (Johnson & Neyman, 1936). Dieses Intervall trennt Bereiche auf der Moderatorvariablen voneinander ab, in denen ein bedingtes Regressionsgewicht signifikant bzw. nicht signifikant von 0 verschieden ist. In unserem Anwendungsbeispiel weist das Johnson-Neyman-Intervall, das zu einem Signifikanzniveau von $\alpha = 0,05$ korrespondiert, folgende Intervallgrenzen auf: $JNI = [-90,451; -0,260]$. Außerhalb dieses Bereichs sind die bedingten Regressionsgewichte von 0 verschieden. Da keine Extraversionswerte kleiner als $-90,451$ im Datensatz vorkommen, ist nur die rechte Grenze des Intervalls interessant. Für Extraversionswerte größer als $-0,260$ gibt es einen bedeutsamen Zusammenhang zwischen emotionaler Ansteckung und dem Wohlbefinden. Das bedingte Regressionsgewicht an der Intervallgrenze beträgt $b_{-0,260} = 0,138$. Die emotionale Ansteckung setzt somit eine bestimmte Extraversion voraus, um sich positiv auf das Wohlbefinden auszuwirken. In unserem Beispiel sind die bedingten Regressionsgewichte für Extraversionswerte, die außerhalb des Intervalls liegen, signifikant von 0 verschieden. Es kann aber auch Anwendungen geben, für die es umgekehrt ist, d. h., für die die bedingten Regressionsgewichte bedeutsam von 0 verschieden sind für Werte der Moderatorvariablen, die innerhalb des Intervalls liegen. Die entsprechenden Informationen müssen der Ausgabe des Computerprogramms entnommen werden. Johnson-Neyman-Intervalle lassen sich mit dem Internet-Tool von Preacher et al. (2006) sowie mit spezifischen R-Paketen (Luhmann, 2013) bestimmen.

Konfidenzbänder. Bereiche der Moderatorvariablen, in denen die zugeordneten bedingten Regressionsgewichte signifikant von 0 verschieden sind, lassen sich auch anhand von Konfidenzbändern bestimmen. Hierzu be-

stimmt man für alle zulässigen Werte der Moderatorvariablen Konfidenzintervalle nach Gleichung F 19.72. Diese kann man auch grafisch darstellen (s. Abb. 19.8). Die bedingten Regressionsgewichte sind für Werte der Moderatorvariablen (Extraversion), bei denen die Konfidenzbänder die 0 ausschließen, signifikant von 0 verschieden.

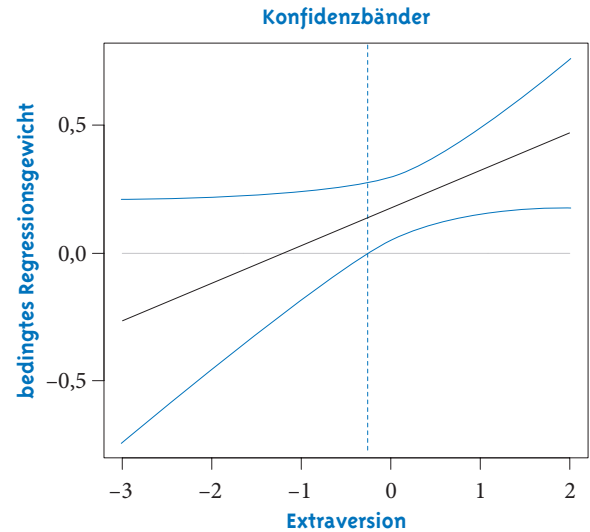


Abbildung 19.8 95 %-Konfidenzbänder für die bedingten Regressionsgewichte (simple slopes), die den Einfluss der emotionalen Ansteckung auf das Wohlbefinden repräsentieren, in Abhängigkeit von der Moderatorvariablen Extraversion

Auswirkungen des Messfehlers

Der Regressionsanalyse liegt die Annahme zugrunde, dass die unabhängigen Variablen messfehlerfrei gemessen wurden. Wie wir in Kapitel 23 sehen werden, ist diese Annahme für viele Fragestellungen der Psychologie problematisch, da man Messfehlereinflüsse bei vielen Messungen nicht vermeiden kann. Dies ist insbesondere dann der Fall, wenn man von stochastischen Regressoren ausgeht. Messfehler in den unabhängigen Variablen führen dazu, dass die Regressionsgewichte nicht mehr erwartungstreu, nicht konsistent und mit geringer Effizienz geschätzt werden. Dies ist besonders bei moderierten Regressionsanalysen gravierend, da die Produktvariablen messfehlerbehafteter sind als die unabhängigen Variablen, auf deren Grundlage sie berechnet wurden. Daraus folgt, dass Moderatoreffekte unterschätzt werden und die Teststärke sich verringert. Je messfehlerbehafteter unabhängige Vari-

ablen sind, desto schwieriger ist es, Moderatoreffekte aufzudecken (Klein, 2000). Im Falle messfehlerbehafteter unabhängiger Variablen sollte man daher zur Analyse von Moderatoreffekten auf Modelle mit latenten Variablen zurückgreifen (s. Kap. 26). Bei diesen Verfahren weist die Zentrierung der Variablen nicht nur Vorteile in Bezug auf die Interpretierbarkeit der Regressionsgewichte auf, sondern auch rechentechnische Vorteile bei der Schätzung der Parameter (Dalal & Zickar, 2012). Bei der Analyse von Moderatorvariablen im Rahmen von linearen Strukturgleichungsmodellen ist daher die Zentrierung der Variablen generell zu empfehlen.

Stochastische Regressoren

Im Fall stochastischer Regressoren trifft man die Annahme, dass die abhängige und die unabhängigen Variablen multivariat normalverteilt sind. Sind die unabhängigen Variablen normalverteilt und liegt eine Interaktion zwischen ihnen vor, kann die abhängige Variable nicht mehr normalverteilt sein, sondern folgt einer komplexen Verteilungsform (Klein, 2000). Eine Voraussetzung der inferenzstatistischen Absicherung des Regressionsmodells mit stochastischen Regressoren ist somit im Falle einer Interaktion a priori nicht erfüllt. Im Falle stochastischer Regressoren bietet es sich an, auf Verfahren der Parameterschätzung und der inferenzstatistischen Testung zurückzugreifen, die für lineare Strukturgleichungsmodelle entwickelt wurden.

19.10 Analyse nicht-linearer Zusammenhänge

In Kapitel 17 haben wir die einfache lineare Regression kennengelernt, in der von einem linearen Zusammenhang zwischen zwei Variablen – wie bei der Produkt-Moment-Korrelation – ausgegangen wird. Diese Annahme wurde in diesem Kapitel auf mehrere Variablen übertragen. Der multiplen Regressionsanalyse liegt in der in Formel F 19.1 dargestellten Form die Annahme zugrunde, dass die Regression einer abhängigen auf eine unabhängige Variable linear ist, wenn die anderen Variablen statistisch kontrolliert werden. Dies ist für den Fall zweier unabhängiger Variablen in Abbildung 19.4 (in Abschn. 19.4.1) dargestellt. Die Annahme der Linearität ist für manche Fragestellungen der Psychologie jedoch keine theoretisch angemessene Annahme.

Beispiel

Zufriedenheit mit einem Kurs und wahrgenommene Anforderung

Ein Beispiel, in dem eine lineare Beziehung zwischen zwei Variablen theoretisch wenig sinnvoll ist, ist der Zusammenhang zwischen der Zufriedenheit mit einem Kurs (Seminar) an der Universität und den wahrgenommenen Anforderungen. Eine lineare Beziehung würde bedeuten, dass die Zufriedenheit mit den wahrgenommenen Anforderungen entweder linear anwächst oder linear abfällt. Beide Annahmen sind wenig sinnvoll. Vielmehr weiß man aus der Lehrevaluationsforschung, dass Studierende dann besonders zufrieden sind, wenn ein Kurs eine mittlere Schwierigkeit aufweist. Ist der Kurs zu leicht, ist man schnell unterfordert, langweilt sich und ist unzufrieden. Ist der Kurs hingegen zu schwer, ist man überfordert, frustriert und ebenfalls unzufrieden. Man würde daher einen nicht-linearen, genauer gesagt einen umgekehrt U-förmigen Zusammenhang zwischen beiden Variablen erwarten.

Wie lassen sich nicht-lineare Zusammenhänge modellieren?

Dies geht im Rahmen der multiplen Regressionsanalyse relativ einfach. In Abbildung 19.9 sind drei verschiedene nicht-lineare Formen des Zusammenhangs zwischen einer abhängigen Variablen Y und einer unabhängigen Variablen X angegeben. In den beiden ersten Fällen besteht ein Zusammenhang in Form einer Parabel zwischen den vorhergesagten \hat{y} -Werten und den x -Werten. Man spricht in diesen beiden Fällen auch von einem *quadratischen* Zusammenhang. Im dritten Beispiel ist der Zusammenhang *kubischer* Natur.

Ein nicht-linearer Zusammenhang lässt sich im Rahmen der multiplen Regressionsanalyse beschreiben, indem Polynome höherer Ordnung definiert und in die Analyse einbezogen werden. Für den einfachen Fall einer quadratischen nicht-linearen Abhängigkeit zwischen einer abhängigen Variablen Y und einer unabhängigen Variablen X erhält man die Regressionsgleichung, indem man die x -Werte quadriert und die so entstandene Variable X^2 zusätzlich zu der unquadratierten Variablen X in die Regressionsgleichung mit aufnimmt:

$$\hat{y}_m = b_0 + b_1 \cdot x_m + b_2 \cdot x_m^2 \quad (\text{F } 19.73)$$

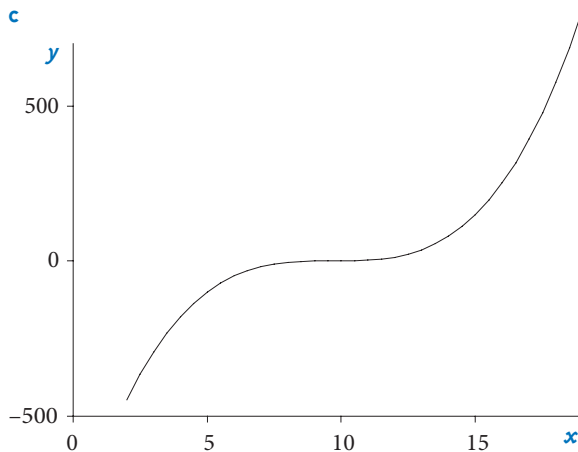
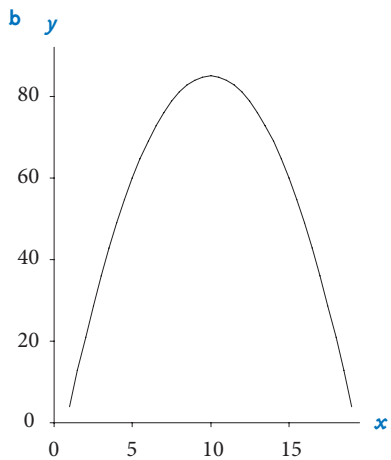
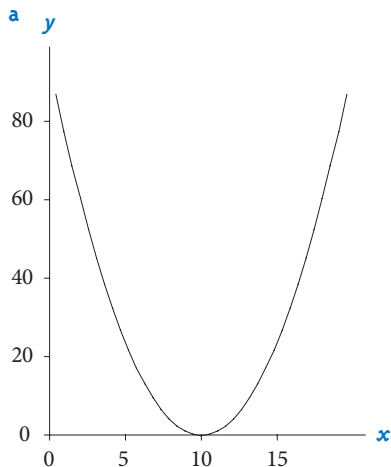


Abbildung 19.9 Verschiedene Formen nicht-linearer Zusammenhänge zwischen zwei Variablen

a Quadratischer Zusammenhang: $y_m = (x_m - 10)^2$

b Quadratischer Zusammenhang: $y_m = 85 - (x_m - 10)^2$

c Kubischer Zusammenhang: $y_m = (x_m - 10)^2 + (x_m - 10)^3$

bzw.

$$y_m = b_0 + b_1 \cdot x_m + b_2 \cdot x_m^2 + e_m \quad (\text{F 19.74})$$

Diese kurvilineare Regression ist also nichts anderes als eine multiple Regressionsanalyse mit den beiden Prädiktoren X und X^2 . Einen kubischen Zusammenhang erhält man, indem man für jeden Wert x_m den Wert x_m^3 berechnet und diese Variable X^3 zusätzlich zu der quadrierten Variablen X^2 und der Variablen X in die Regressionsgleichung aufnimmt:

$$y_m = b_0 + b_1 \cdot x_m + b_2 \cdot x_m^2 + b_3 \cdot x_m^3 + e_m \quad (\text{F 19.75})$$

Dies kann nun beliebig weitergeführt werden, indem sukzessive die Werte x_m^4, x_m^5, \dots in die Gleichung aufgenommen werden. In der Psychologie gibt es aber selten Fälle, in denen ein Zusammenhang vorliegt, der erst mit einem Polynom vierter oder noch höherer Ordnung beschrieben werden kann. Der am häufigsten vorkommende nicht-lineare Zusammenhang ist sicherlich der quadratische Zusammenhang.

Um die quadratische Regression zu illustrieren, haben wir Daten simuliert, die den Zusammenhang zwischen der Zufriedenheit und der wahrgenommenen Anforderung illustrieren, so wie man sie ungefähr in empirischen Studien gewonnen hat. Eine Analyse dieser simulierten Daten mit der multiplen Regressionsanalyse ergab für das Beispiel des Zusammenhangs zwischen Zufriedenheit und Anforderung folgende Regressionsgleichung:

$$\hat{y}_m = 1,433 + 2,641 \cdot x_m - 0,429 \cdot x_m^2$$

In dieser Regressionsgleichung wird der quadratische Term genauso behandelt wie eine andere unabhängige Variable. Für unseren Anwendungsfall ist der Zusammenhang in Abbildung 19.9 a dargestellt. Diese Abbildung enthält auch die einfache lineare Regression, die man erhält, wenn man die quadratische Komponente nicht berücksichtigt. Die einfache lineare Regressionsgleichung lautet:

$$\hat{y}_m = 5,010 + 0,032 \cdot x_m$$

Diese Gleichung und Abbildung 19.10 a zeigen, dass beide Variablen nicht linear-regressiv voneinander abhängig sind. Die Produkt-Moment-Korrelation beträgt $r = 0,033$; die multiple Korrelation, die den quadratischen Zusammenhang berücksichtigt, beträgt hingegen $R = 0,637$ und ist sehr hoch. Dies zeigt, dass man den Zusammenhang zwischen beiden Variablen erheblich

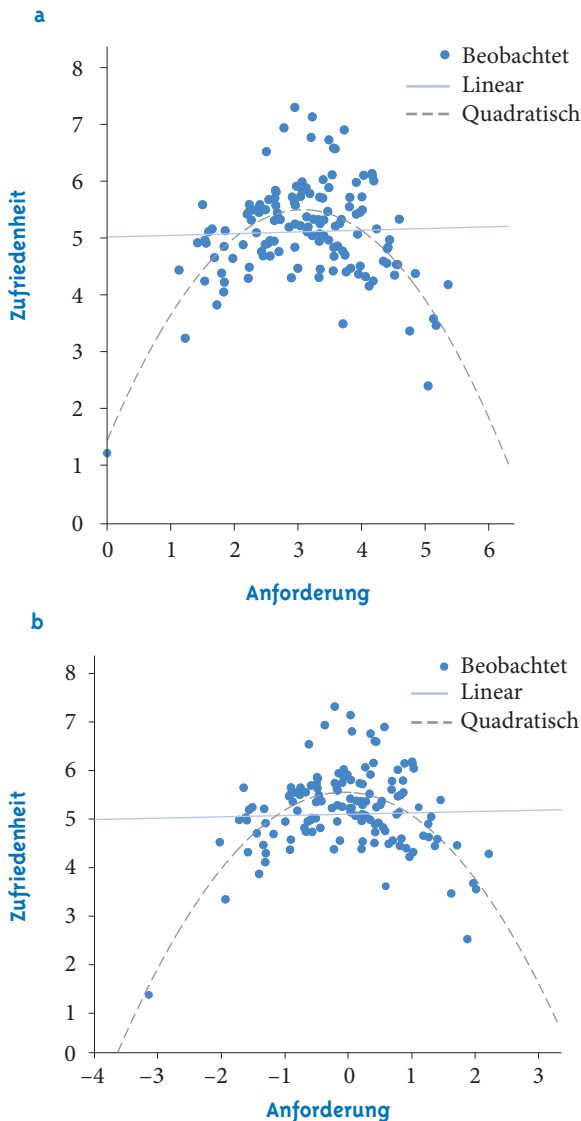


Abbildung 19.10 Ergebnis einer linearen und einer quadratischen Regression für das Beispiel des Zusammenhangs zwischen der Zufriedenheit mit einem Kurs und den wahrgenommenen Anforderungen **a** Unzentrierte Variablen **b** Zentrierte Variablen

verschätzen würde, würde man die quadratische Komponente nicht in die Gleichung mit aufnehmen.

Da bei einer Vielzahl von unabhängigen Variablen die Überprüfung nicht-linearer Effekte sehr aufwendig sein kann, geht man in der Forschungspraxis meist so vor, dass man nicht-lineare Effekte in multiplen Regressionsanalysen nur dann untersucht, wenn es hierfür auch theoretische Argumente gibt.

Zentrierte unabhängige Variablen

Um die Interpretation der Regressionskoeffizienten zu erleichtern, kann man wie bei der moderierten Regression zur Zentrierung greifen. Die Konsequenzen der Zentrierung entsprechen denjenigen bei der moderierten Regression. Dies leuchtet unmittelbar ein, wenn man sich vergegenwärtigt, dass die quadratische Variable X^2 ja nichts anderes ist als eine Produktvariable ($X \cdot X$), wie man sie auch bei der moderierten Regression betrachtet. Die Zentrierung ändert nicht die Form der Beziehung, wie man sich anhand eines Vergleichs der Abbildungen 19.9 a (unzentrierte Werte) und 19.9 b (zentrierte Werte) veranschaulichen kann. Auch ändert sich das geschätzte Regressionsgewicht der quadratischen Komponente nicht, wie man der Regressionsgleichung entnehmen kann:

$$\hat{y}_m = 5,497 - 0,053 \cdot x_m - 0,429 \cdot x_m^2$$

Die Zentrierung erleichtert die Interpretation des Achsenabschnitts, falls der Wert 0 der unabhängigen Variablen X keine klare Bedeutung hat oder in den Daten nicht vorkommt. Der Achsenabschnitt bei zentrierten Variablen entspricht dem y -Wert, den man für Personen erwartet, die eine mittlere Ausprägung auf X aufweisen. In unserem Anwendungsbeispiel erwartet man für Personen, die die Anforderungen des Kurses so wahrnehmen, dass sie genau dem Mittelwert der Stichprobe entsprechen, einen Zufriedenheitswert von 5,497.

Inferenzstatistische Absicherung

Zur Überprüfung der Hypothese, dass ein nicht-linearer Zusammenhang zwischen der abhängigen und der unabhängigen Variablen vorliegt, geht man wie bei anderen Modellen der multiplen Regressionsanalyse vor. Man überprüft die Hypothese, dass das Regressionsgewicht der nicht-linearen Komponente, die von Interesse ist, in der Population nicht von 0 verschieden ist. Wie bei der moderierten Regressionsanalyse ist es wichtig, dass alle Terme einer nicht-linearen Gleichung in die Regressionsgleichung mit aufgenommen werden. Überprüft man z. B. ein quadratisches Regressionsmodell, müssen neben den x^2 -Werten auch die x -Werte in die Regressionsgleichung aufgenommen werden. Dies ist notwendig, da die quadrierte Variable X^2 mit der Ausgangsvariablen X hoch korreliert ist und somit der Einfluss der Variablen selbst kontrolliert werden muss. Würde man die Ausgangsvariable nicht mit in die Regressionsgleichung aufnehmen, so könnte ein bedeutsamer Einfluss des quadratischen

Terms in Wirklichkeit bloß einen einfachen linearen Einfluss widerspiegeln, sodass die Schlussfolgerung, dass es sich um eine nicht-lineare Abhängigkeit handelt, falsch wäre. Bezogen auf unser Beispiel zeigt sich, dass im Modell mit unzentrierten Variablen alle Regressionskoeffizienten bedeutsam von 0 verschieden sind ($\alpha = 0,05$), im Modell mit zentrierten Variablen sind nur der Achsenabschnitt und das Regressionsgewicht der quadratischen Komponente signifikant von 0 verschieden.

19.11 Analyse kategorialer unabhängiger Variablen

In den bisherigen Anwendungen der multiplen Regressionsanalyse haben wir metrische unabhängige Variablen analysiert. Die multiple Regressionsanalyse kann auch herangezogen werden, um den Einfluss kategorialer unabhängiger Variablen zu überprüfen. Das klassische Verfahren zur Analyse einer metrischen abhängigen Variablen und einer kategorialen unabhängigen Variablen ist die einfaktorielle Varianzanalyse, die wir bereits in Kapitel 13 behandelt haben. Wie Cohen (1968) gezeigt hat, handelt es sich bei der multiplen Regressionsanalyse und der Varianzanalyse um zwei Spezialfälle eines allgemeinen statistischen Modells zur Analyse metrischer abhängiger Variablen, das Allgemeine Lineare Modell (ALM) genannt wird. Um kategoriale unabhängige Variablen im Rahmen der Regressionsanalyse untersuchen zu können, ist es notwendig, die Information, die in den unabhängigen kategorialen Variablen enthalten ist, zunächst anhand von Codiervariablen abzubilden. Wir werden im Folgenden zwei Möglichkeiten der Bildung von Codiervariablen vorstellen: zum einen die Dummy-Codierung und zum anderen die Effektcodierung. In beiden Fällen werden zur Codierung einer Variablen, die aus c Kategorien besteht, nur $c - 1$ Codiervariablen benötigt.

! Im Allgemeinen werden zur Codierung einer unabhängigen Variablen, die aus c Kategorien besteht, $c - 1$ Codiervariablen benötigt.

19.11.1 Dummy-Codierung

Wir wollen die Grundidee der Dummy-Codierung anhand eines konkreten empirischen Beispiels einführen.

Einführendes Beispiel: Materielle Situation und Stadtgröße

Die abhängige Variable in unserem Beispiel ist die materielle Situation einer Person. Hierbei handelt es sich um den Mittelwert von z -transformierten Einzelindikatoren, die unterschiedliche Ausgangsskalierungen aufwiesen. Diese Einzelindikatoren erfassen u. a. die berufliche Situation einer Person (z. B. Stellenumfang, Befristungsstatus, hierarchische Position), ihr Einkommen, ihre Ersparnisse und andere Vermögenswerte wie Immobilien und Wertgegenstände sowie ihre Wohnsituation (Eigentumsverhältnisse, Größe und Ausstattungsniveau der Wohnung oder des Hauses). Wir wollen nun untersuchen, inwieweit die materielle Situation einer Person davon abhängt, ob sie in einer Großstadt, Mittelstadt oder Kleinstadt wohnt. Die unabhängige Variable ist somit der »Größenstatus« einer Stadt, eine kategoriale Variable, die drei Kategorien aufweist. Die Daten entstammen einer Untersuchung von Schmitt und Maes (2001), in der die Lebensqualität von Deutschen untersucht wurde. Hierzu wurden sowohl in den neuen als auch in den alten Bundesländern Personen befragt, die in Klein-, Mittel- oder Großstädten leben. Der Untersuchungsplan war so angelegt, dass für alle sechs Zellen (zwei Regionen mal drei Stadtgrößen) dieselbe Anzahl von Personen untersucht werden sollten. Aufgrund von Untersuchungsausfällen sind jedoch nicht alle sechs Zellen gleich häufig besetzt.

In einem ersten Schritt wollen wir anhand der kategorialen Variablen Stadtgröße zeigen, wie kategoriale Variablen codiert werden können. In einem späteren Schritt werden wir dann die Region (West – Ost) hinzunehmen. Wie sehen die Daten aus? In Tabelle 19.15 a sind die Werte von jeweils vier Personen, die in einer Großstadt, einer Mittelstadt bzw. einer Kleinstadt leben, dargestellt. Tabelle 19.15 b enthält die codierte Datenmatrix. Da die unabhängige Variable Stadtgröße drei Kategorien aufweist, werden zwei Codiervariablen X_1 und X_2 benötigt. Die erste Spalte der Datenmatrix in Tabelle 19.15 b enthält die Werte der 12 Personen auf der abhängigen Variablen (materielle Situation). Die Werte auf der ersten Codiervariablen (X_1) resultieren daraus, dass Personen, die in einer Großstadt leben, den Wert 1 zugeordnet bekommen, während alle anderen Personen den Wert 0 erhalten. Die erste Codiervariable zeigt somit an, ob eine Person in einer Großstadt lebt oder nicht. Die Codierung auf

der zweiten Codiervariablen (X_2) wird so vorgenommen, dass alle Personen, die in einer Mittelstadt leben, den Wert 1 zugeordnet bekommen, während alle anderen Personen auf dieser Codiervariablen den Wert 0 erhalten. Die zweite Codiervariable zeigt somit an, ob eine Person in einer Mittelstadt lebt oder nicht. Personen, die in einer Kleinstadt leben, erhalten auf beiden Codiervariablen eine 0. Somit lässt sich anhand beider Codiervariablen auch eindeutig erkennen, ob eine Person in einer Kleinstadt lebt. Eine dritte Codiervariable, die die Zugehörigkeit zur Kleinstadt derart anzeigen würde, dass alle Personen, die in einer Kleinstadt leben, eine 1 und alle anderen Personen eine 0 zugeordnet bekommen, wäre somit redundant. Durch beide Codiervariablen wird also die gesamte Information codiert, die in der Zugehörigkeit zu einer Stadt einer bestimmten Größe enthalten ist. (Dies zeigt sich darin, dass Personen unterschiedlicher Stadtgrößen eine unterschiedliche Wertekombination auf beiden Codiervariablen aufweisen.) Die in unserem Beispiel vorgenommene Codierung illustriert die allgemeine Regel der Dummy-Codierung, die wir im Folgenden zusammenfassen.

Definition

Die **Dummy-Codierung** erfolgt in folgenden Schritten:

- (1) Wähle eine der c Kategorien der unabhängigen Variablen als Referenzkategorie aus.
- (2) Weise dieser Referenzkategorie auf allen Codiervariablen den Wert 0 zu.
- (3) Weise allen anderen Kategorien der unabhängigen Variablen Werte auf den Codiervariablen derart zu, dass
 - (a) jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0,
 - (b) jede Codiervariable nur für eine einzige Kategorie den Wert 1 aufweist, für alle anderen den Wert 0.

Die codierte Datenmatrix kann einer multiplen Regressionsanalyse unterzogen werden, in die die abhängige Variable Y und die beiden unabhängigen Variablen X_1 und X_2 eingehen. Einige Statistikprogramme übernehmen die Bildung der Dummy-Variablen, sodass diese nicht von dem Anwender gebildet werden müssen. Bei dem Statistikprogramm SPSS entspricht die Dummy-

Codierung z.B. dem »einfachen Kontrast« unter der Option »Kontrastbildung« beim Allgemeinen Linearen Modell. Die multiple Regressionsanalyse ergibt für unser Datenbeispiel folgende Gleichung:

$$\hat{Y} = -0,054 + 0,132 \cdot X_1 + 0,001 \cdot X_2 \quad (\text{F } 19,76)$$

Die Regressionskoeffizienten haben bei einer Dummy-Codierung eine besondere Bedeutung.

Tabelle 19.15 Darstellung einer Dummy-Codierung anhand einer metrischen abhängigen Variablen (materielle Situation) und einer kategorialen unabhängigen Variablen (Größenstatus einer Stadt: Groß-, Mittel- oder Kleinstadt)

a Auszug aus den Rohdaten: Werte der abhängigen Variablen (materielle Situation) von jeweils vier Personen einer Großstadt, einer Mittelstadt und einer Kleinstadt

Großstadt	Mittelstadt	Kleinstadt
0,41	−0,36	−0,25
−0,25	−0,32	−0,44
0,00	−0,52	−0,15
0,10	0,13	0,16

b Auszug aus der dummy-codierten Datenmatrix

Y	X_1	X_2	
0,41	1	0	Großstadt
−0,25	1	0	
0,00	1	0	
0,10	1	0	
−0,36	0	1	Mittelstadt
−0,32	0	1	
−0,52	0	1	
0,13	0	1	
−0,25	0	0	Kleinstadt
−0,44	0	0	
−0,15	0	0	
0,16	0	0	

Bedeutung des Achsenabschnitts b_0

Die Regressionskonstante b_0 entspricht dem Mittelwert der abhängigen Variablen in der Referenzkategorie, d.h. der Personengruppe, die auf allen Codiervariablen eine 0 aufweist. Dass dem Achsenabschnitt b_0 diese Bedeutung zukommt, wird deutlich, wenn man in die Gleichung F 19.76 jeweils den Wert 0 für die Codiervariablen X_1 und X_2 einsetzt. Für die Wertekombination $x_1 = 0$ und $x_2 = 0$ ergibt sich dann der erwartete y -Wert von $\hat{y} = -0,054$. Dies ist genau der Mittelwert der Personen, die in einer Kleinstadt leben, da die Kleinstadt als Referenzgruppe gewählt wurde. Der \hat{y} -Wert ist gleich dem Mittelwert in dieser Kategorie, da der Mittelwert der beste Schätzwert nach dem Kleinste-Quadrate-Kriterium ist. Wie man Tabelle 19.16 entnehmen kann, entspricht b_0 genau dem Mittelwert in der Kleinstadt-Bedingung.

Tabelle 19.16 Mittelwerte der abhängigen Variablen in den verschiedenen Größenstatus-Kategorien

Kategorie	Mittelwert
Großstadt ($n = 896$)	0,078
Mittelstadt ($n = 755$)	-0,053
Kleinstadt ($n = 733$)	-0,054
Gesamtmittelwert (gewichteter Mittelwert)	-0,004
Ungewichteter Mittelwert der Mittelwerte in den Kategorien	-0,010

Bedeutung der Regressionsgewichte b_j

Das Regressionsgewicht b_1 der ersten Dummy-Variablen entspricht der Differenz zwischen dem Mittelwert der Kategorie, die auf der Variablen X_1 eine 1 zugeordnet bekommen hat, und dem Mittelwert der Referenzgruppe. In unserem Beispiel entspricht also b_1 der Differenz zwischen dem Mittelwert der Personen, die in einer Großstadt leben, und dem Mittelwert der Personen, die in einer Kleinstadt leben. Auch das lässt sich zeigen, indem in Gleichung F 19.76 für X_1 und für X_2 diejenigen Werte eingesetzt werden, die bei der Dummy-Codierung für die Kategorie »Großstadt« ($x_1 = 1$ und $x_2 = 0$) festgelegt wurden. Hieraus folgt: $\hat{y} = -0,054 + 0,132 = 0,078$. In analoger Weise entspricht das Regressionsgewicht b_2 der zweiten Dummy-Variablen der Differenz zwischen dem Mittelwert der

Gruppe, die auf der zweiten Dummy-Variablen eine 1 zugeordnet bekommen hat, und der Referenzgruppe. In unserem Beispiel ist dies die Differenz zwischen dem Mittelwert der Personen, die in einer Mittelstadt leben, und dem Mittelwert der Personen, die in einer Kleinstadt leben.

Aus diesem Beispiel wird deutlich, dass die Codiervariablen bei der Dummy-Codierung dazu dienen, die Mittelwerte der abhängigen Variablen zwischen zwei Kategorien der unabhängigen Variablen zu vergleichen. Es handelt sich dabei immer um den Vergleich einer Referenzkategorie mit einer anderen der $c - 1$ Kategorien. Die Regressionsgewichte der Codiervariablen geben Richtung und Größe des jeweiligen Unterschieds zwischen den verglichenen Mittelwerten an. Wir haben im Rahmen der Varianzanalyse solche paarweisen Unterschiede zwischen Kategorien einer unabhängigen Variablen in Bezug auf die abhängige Variable als Kontraste bezeichnet. Eine Dummy-Codierung erlaubt also die Modellierung (und Testung) einfacher Kontraste zwischen Kategorienpaaren.

In unserem Beispiel gibt es neben dem Unterschied zwischen Klein- und Großstadt und dem Unterschied zwischen Klein- und Mittelstadt noch einen dritten Unterschied, nämlich den zwischen Mittel- und Großstadt. Dieser Unterschied wurde nicht über eine eigene Codiervariable modelliert. Das war auch nicht nötig, denn der Unterschied zwischen Mittel- und Großstadt lässt sich direkt aus den beiden anderen Kontrasten berechnen:

$$\begin{aligned} & \text{Großstadt} - \text{Mittelstadt} \\ &= (\text{Großstadt} - \text{Kleinstadt}) \\ &\quad - (\text{Mittelstadt} - \text{Kleinstadt}) \end{aligned}$$

Die Richtung und Größe des Unterschieds zwischen Großstadt und Mittelstadt ist also nichts anderes als die Differenz zwischen den Regressionsgewichten b_1 und b_2 . In unserem Beispiel ergibt sich für diese Differenz $b_1 - b_2 = 0,132 - 0,001 = 0,131$. Dies ist der Unterschied zwischen dem Mittelwert der Personen, die in einer Mittelstadt leben, und dem Mittelwert der Personen, die in einer Großstadt leben. Das lässt sich wiederum an den in Tabelle 19.16 dargestellten Mittelwerten leicht nachvollziehen. Der dritte Kontrast lässt sich also aus den ersten beiden Kontrasten bestimmen. Er wird daher in die multiple Regressionsanalyse auch nicht aufgenommen. Das bedeutet allerdings auch, dass der dritte Kontrast nicht direkt inferenzstatistisch

abgesichert werden kann. Ist man aus theoretischen Gründen an einer inferenzstatistischen Absicherung des Kontrasts zwischen Mittel- und Großstadt in Bezug auf die finanzielle Situation der Einwohner interessiert, hätte man die Dummy-Codierung anders vornehmen müssen, etwa indem man die Kategorie »Mittelstadt« zur Referenzkategorie macht.

19.11.2 Effektcodierung

Eine weitere Möglichkeit der Codierung kategorialer Variablen ist die Effektcodierung. Bei der Effektcodierung unterscheidet man zwei Formen: die ungewichtete und die gewichtete Effektcodierung. Wir werden beide Formen zunächst vorstellen und illustrieren, um danach zu diskutieren, unter welchen Bedingungen man zu welcher der beiden Formen greift.

Ungewichtete Effektcodierung

Eine ungewichtete Effektcodierung der in Tabelle 19.15 a angegebenen Daten ist in Tabelle 19.17 dargestellt. Wie bei der Dummy-Codierung werden auch bei der Effektcodierung nur zwei Codiervariablen benötigt, um die gesamte Information, die in den Daten steckt, zu codieren. Bei der Effektcodierung geht man wie folgt vor. Man wählt wieder eine Referenzkategorie aus. Diese Referenzkategorie erhält nun auf beiden Codiervariablen nicht eine 0, sondern eine -1 . Ansonsten unterscheiden sich die Werte der Codiervariablen nicht von denen der Dummy-Codierung, wie man Tabelle 19.17 entnehmen kann.

Tabelle 19.17 Codierte Datenmatrix für eine ungewichtete Effektcodierung der Daten in Tabelle 19.15 a

Y	X ₁	X ₂	
0,41	1	0	Großstadt
-0,25	1	0	
0,00	1	0	
0,10	1	0	
-0,36	0	1	Mittelstadt
-0,32	0	1	
-0,52	0	1	
0,13	0	1	

Tabelle 19.17 (Fortsetzung)

Y	X ₁	X ₂	
-0,25	-1	-1	Kleinstadt
-0,44	-1	-1	
-0,15	-1	-1	
0,16	-1	-1	

Definition

Die **allgemeine Regel der ungewichteten Effektcodierung** lautet:

- (1) Wähle eine der c Kategorien der unabhängigen Variablen als Referenzkategorie aus.
- (2) Weise der Referenzkategorie auf allen Codiervariablen den Wert -1 zu.
- (3) Weise allen anderen Kategorien der unabhängigen Variablen einen Wert auf den $c - 1$ Codiervariablen derart zu, dass
 - (a) jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0,
 - (b) jede Codiervariable nur für eine einzige Kategorie den Wert 1 und für die Referenzkategorie den Wert -1 aufweist, für alle anderen Kategorien den Wert 0.

Angewendet auf unser Beispiel bedeutet dies, dass Personen, die in einer Großstadt leben, auf der ersten Codiervariablen eine 1 und auf der zweiten Codiervariablen eine 0 zugewiesen bekommen. Personen, die in der Mittelstadt leben, erhalten auf der ersten Codiervariablen eine 0 und auf der zweiten Codiervariablen eine 1. Personen die in der Kleinstadt leben, bekommen auf beiden Codiervariablen den Wert -1 zugeordnet. Das Statistikprogramm SPSS nimmt diese Zuordnung von Werten auf den Codiervariablen in der Kontrast-Option »Abweichung« beim Allgemeinen Linearen Modell selbstständig vor. Eine multiple Regressionsanalyse der Daten unseres Beispiels mit diesen Codiervariablen ergibt:

$$\hat{Y} = -0,010 + 0,088 \cdot X_1 - 0,043 \cdot X_2$$

Bei der ungewichteten Effektcodierung haben die Regressionsparameter eine andere Bedeutung als bei der Dummy-Codierung und als bei der gewichteten Effektcodierung.

Bedeutung des Achsenabschnitts b_0 . Die Regressionskonstante b_0 entspricht dem *ungewichteten* Mittelwert der Mittelwerte in den drei Kategorien. Dies bedeutet, dass die drei Mittelwerte der Großstadt, Mittelstadt und Kleinstadt addiert und durch 3 geteilt werden. Der ungewichtete Mittelwert berücksichtigt nicht, dass die drei Kategorien der unabhängigen Variablen unterschiedliche Größen haben können. Dies hat zur Folge, dass sich der ungewichtete Mittelwert von dem Gesamtmittelwert über alle Personen unterscheiden kann.

In unserem Datenbeispiel sind die Stichprobengrößen in den drei Kategorien der Stadtgröße relativ unterschiedlich. Von daher ist es nicht verwunderlich, dass der Gesamtmittelwert ($-0,004$) von dem Durchschnitt der drei Kategorienmittelwerte ($-0,010$) abweicht (s. Tab. 19.16). Der Achsenabschnitt b_0 entspricht also nur dann dem Gesamtmittelwert, wenn alle Teilstichproben gleich groß sind.

Bedeutung der Regressionsgewichte b_j . Das Regressionsgewicht b_1 der ersten Codiervariablen entspricht der Differenz zwischen dem Mittelwert der Kategorie, die auf der Variablen X_1 eine 1 zugeordnet bekommen hat, und dem ungewichteten Mittelwert über alle Kategorien hinweg (also dem Wert, der auch durch den Achsenabschnitt b_0 ausgedrückt wird). In unserem Beispiel zeigt er somit an, dass Bewohner einer Großstadt im Mittel um den Wert 0,088 wohlhabender sind als der Durchschnitt. Der Regressionsparameter b_2 entspricht der Differenz zwischen dem Mittelwert der Kategorie, die auf der Variablen X_2 eine 1 zugeordnet bekommen hat, und dem ungewichteten Mittelwert über alle Kategorien hinweg. In unserem Beispiel zeigt b_2 an, dass Bewohner einer Mittelstadt im Mittel um den Wert 0,043 weniger wohlhabend sind als der Durchschnitt. Der Mittelwert der Kleinstadt (Referenzkategorie) ist durch keinen Parameter repräsentiert, sondern muss aus den anderen berechnet werden:

$$\hat{y}_{\text{Kleinstadt}} = b_0 - b_1 - b_2 = -0,010 - 0,088 + 0,043 = -0,055$$

Der Unterschied zum Mittelwert in Tabelle 19.16 ist auf kleine Unschärfen bei der Rundung zurückzuführen.

Gewichtete Effektcodierung

Bei einer gewichteten Effektcodierung bildet man die Werte der Codiervariablen derart, dass sie nicht Abweichungen vom ungewichteten Mittelwert der Ka-

tegorien, sondern Abweichungen vom gewichteten Mittelwert (also dem Gesamtmittelwert) darstellen. Der gewichtete Mittelwert wird gebildet, indem die Mittelwerte in den drei Kategorien mit ihrer relativen Häufigkeit gewichtet und aufsummiert werden. Diesen Kennwert hatten wir in Abschnitt 6.4.2 als gewogenes arithmetisches Mittel (*GAM*) kennengelernt (Formel F 6.16):

$$GAM_X = \frac{\sum_{r=1}^s \bar{x}_r \cdot n_r}{\sum_{r=1}^s n_r}$$

Das *GAM* entspricht dem Gesamtmittelwert, d. h. dem Mittelwert, den wir erhalten, wenn wir die individuellen Werte aufsummieren und durch n teilen.

Um also zu erreichen, dass der Achsenabschnitt b_0 dem Gesamtmittelwert entspricht und die Regressionsgewichte b_j spezifischen Abweichungen von diesem Gesamtmittelwert entsprechen, muss der Wert, der der Referenzkategorie zugeordnet wird, geändert werden. Der Referenzkategorie wird nicht mehr der Wert von -1 auf allen Codiervariablen zugewiesen, sondern auf jeder Codiervariablen X_j ein anderer Wert. Auf der ersten Codiervariablen X_1 wird der Referenzkategorie der Wert $-(n_{X_1}/n_R)$ zugewiesen, wobei mit n_{X_1} die Stichprobengröße derjenigen Kategorie bezeichnet wird, der auf der Codiervariablen X_1 eine 1 zugewiesen wurde (in unserem Beispiel entspricht das der Anzahl der Personen, die in einer Großstadt leben, also $n = 896$), während mit n_R die Stichprobengröße in der Referenzkategorie bezeichnet wird (in unserem Beispiel die Anzahl der Personen, die in einer Kleinstadt leben, also $n = 733$). In unserem Beispiel wird der Referenzkategorie auf der ersten Codiervariablen X_1 also der Wert $-(896/733) = -1,222$ zugewiesen. Auf der zweiten Codiervariablen X_2 wird der Referenzkategorie der Wert $-(n_{X_2}/n_R)$ zugewiesen, wobei mit n_{X_2} die Stichprobengröße derjenigen Kategorie bezeichnet wird, der auf der Codiervariablen X_2 eine 1 zugewiesen wurde (in unserem Beispiel entspricht das der Anzahl aller Personen, die in einer Mittelstadt leben, also $n = 755$). In unserem Beispiel wird der Referenzkategorie auf der zweiten Codiervariablen X_2 also der Wert $-(755/733) = -1,03$ zugewiesen. Allgemein lautet die Formel zur Bestimmung der Werte, die der Referenzkategorie auf einer Codiervariablen X_j im Falle einer gewichteten Effektcodierung zugewiesen werden, demnach $-(n_{X_j}/n_R)$.

Eine gewichtete Effektcodierung wird von Statistikprogrammen im Allgemeinen nicht automatisch erzeugt, sondern muss eigenhändig vorgenommen werden. Die gewichtete Effektcodierung der in Tabelle 19.15 a angegebenen Daten ist in Tabelle 19.18 dargestellt.

Um für unser Beispiel die Regressionskoeffizienten zu bestimmen, verwenden wir die in Tabelle 19.18 dargestellte Effektcodierung. Eine multiple Regressionsanalyse mit diesen Codiervariablen ergibt:

$$\hat{Y} = -0,004 + 0,082 \cdot X_1 - 0,049 \cdot X_2$$

Damit entspricht der Wert von b_0 jetzt genau dem Gesamtmittelwert von $-0,004$ (s. Tab. 19.16). Das Regressionsgewicht b_1 entspricht der Differenz zwischen dem Mittelwert der Großstadt und dem Gesamtmittelwert, das Regressionsgewicht b_2 der Differenz zwischen dem Mittelwert der Mittelstadt und dem Gesamtmittelwert.

Tabelle 19.18 Codierte Datenmatrix für eine gewichtete Effektcodierung der Daten in Tabelle 19.15 a

Y	X ₁	X ₂	
0,41	1	0	Großstadt
-0,25	1	0	
0,00	1	0	
0,10	1	0	
-0,36	0	1	Mittelstadt
-0,32	0	1	
-0,52	0	1	
0,13	0	1	
-0,25	-1,222	-1,030	Kleinstadt
-0,44	-1,222	-1,030	
-0,15	-1,222	-1,030	
0,16	-1,222	-1,030	

Definition

Die **allgemeine Regel der gewichteten Effektcodierung** lautet:

- (1) Wähle eine der c Kategorien der unabhängigen Variablen als Referenzkategorie aus.

- (2) Weise der Referenzkategorie auf der Codiervariablen X_j den Wert $-(n_{X_j}/n_R)$ zu.
- (3) Weise allen anderen Kategorien der unabhängigen Variablen einen Wert auf den $c - 1$ Codiervariablen derart zu, dass
 - (a) jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0,
 - (b) jede Codiervariable nur für eine einzige Kategorie den Wert 1 und für die Referenzkategorie den Wert $-(n_{X_j}/n_R)$ aufweist, für alle anderen Kategorien den Wert 0.

Gewichtete oder ungewichtete Effektcodierung?

Auf die ungewichtete Effektcodierung greift man üblicherweise dann zurück, wenn die Unterschiede in den Größen der Teilstichproben auf unsystematische Ausfälle zurückgeführt werden können. Ein typisches Anwendungsgebiet der ungewichteten Effektcodierung sind experimentelle Studien, in denen sich die Bedingungen aufgrund unsystematischer Unterschiede in ihrer Stichprobengröße unterscheiden, z. B. durch Ausfälle von Versuchspersonen. Letztlich gewichtet man also inhaltlich – und dann auch statistisch – alle Bedingungen gleich. Der Gesamtmittelwert soll den Mittelwert der verschiedenen Bedingungen repräsentieren und nicht von Unterschieden in den Kategorienhäufigkeiten abhängen. Da unserem Beispiel ein solcher Stichprobenplan zugrunde liegt, würde man in unserem Beispiel zur ungewichteten Effektcodierung greifen.

Auf die gewichtete Effektcodierung greift man in Anwendungen zurück, bei denen die Unterschiede in den Kategorienhäufigkeiten inhaltlich relevant sind und berücksichtigt werden müssen. Dies ist insbesondere dann der Fall, wenn die Unterschiede zwischen den Größen der Teilstichproben repräsentativ für die Population sind. Dies wäre in unserem Beispiel der Fall, wenn per Zufall aus der Bevölkerung Deutschlands eine repräsentative Stichprobe gezogen worden wäre und man registriert hätte, ob die Personen in einer Groß-, Mittel- oder Kleinstadt leben. Die Unterschiede in den Kategorienhäufigkeiten kommen dadurch zustande, dass wir eine Zufallsstichprobe gezogen haben und die Unterschiede zwischen den drei Kategorien die Unterschiede in der Population widerspiegeln. Auf die gewichtete Effektcodierung würde man in unserem Beispiel auch dann zurückgreifen, wenn man die drei

Stichproben für die Groß-, Mittel- oder Kleinstädte nachträglich unterschiedlich gewichten wollte, um einen adäquateren Schätzwert für den Gesamtmittelwert Deutschlands zu erhalten.

19.11.3 Vergleich von Dummy- und Effektcodierung

Die unterschiedlichen Bedeutungen der Regressionskoeffizienten bei der Dummy- bzw. Effektcodierung sind in Tabelle 19.19 einander gegenübergestellt. Welche Form der Codierung man wählt, hängt von der Fragestellung einer Untersuchung ab. Lässt sich aufgrund des Versuchsdesigns eine Referenzgruppe klar festlegen, so ist die Dummy-Codierung die Methode der Wahl. Ein typisches Beispiel ist eine experimentelle Studie, in der verschiedene Experimentalbedingungen mit einer Kontrollbedingung verglichen werden. Ein anderes Beispiel ist der Vergleich verschiedener Interventionen mit einer Kontrollbedingung (z.B. Wartekontrollgruppe oder Placebo-Kontrollgruppe). In beiden Fällen kommt der Kontrollbedingung die klare Bedeutung einer Referenzgruppe zu und man ist u.a. daran interessiert, inwieweit sich die einzelnen experimentellen bzw. Interventionsbedingungen von dieser Kontrollbedingung unterscheiden. Für die Auswertung dieser Fragestellungen hat die Dummy-Codierung Vorteile, da die Regressionsgewichte der Dummy-Codierung genau den Kontrast zwischen einzelnen experimentellen bzw. Interventionsgruppen und der Kontrollgruppe abbilden.

Ist die Wahl einer Referenzgruppe jedoch nicht so eindeutig möglich, so bietet sich die Effektcodierung an, wenn man daran interessiert ist, inwieweit sich einzelne Bedingungen von dem Gesamtmittelwert über

alle Bedingungen hinweg unterscheiden. Ein Beispiel ist eine experimentelle Untersuchung, in der unterschiedliche Ausprägungen auf der unabhängigen Variablen miteinander verglichen werden sollen, ohne dass eine Kontrollbedingung realisiert wurde (z.B. wenn man drei Therapieformen hinsichtlich ihrer Effektivität miteinander vergleichen möchte, dabei aber keine Kontrollgruppe realisieren konnte oder wollte). Ein zweites Beispiel ist eine Untersuchung, in der verschiedene Altersgruppen von Jugendlichen (z.B. 12–14 Jahre, 14–16 Jahre, 16–18 Jahre) hinsichtlich einer bestimmten abhängigen Variablen (z.B. Aggressivität) miteinander verglichen werden sollen. Auch hier wird man eher daran interessiert sein, Abweichungen zwischen einer jeweiligen Alterskategorie und dem Gesamtmittelwert über alle Altersgruppen hinweg zu untersuchen, und daher eine Effektcodierung wählen.

Alle vorgestellten Codierformen unterscheiden sich nicht in Bezug auf die aufgeklärte Varianz der abhängigen Variablen. Dies werden wir im Abschnitt 19.11.4 zur Inferenzstatistik zeigen.

Weitere Codierformen

Neben der Dummy- und der Effektcodierung existieren noch andere Codierformen. So haben wir z.B. in Kapitel 13 zur Varianzanalyse lineare Kontraste kennengelernt. Diese Kontraste lassen sich direkt auf die Bestimmung von Codiervariablen übertragen. Wir haben bei der Definition orthogonaler Kontraste in Abschnitt 13.1.13 gesehen, dass bei diesen Kontrasten die Mittelwerte einzelner Bedingungen (Kategorien) gewichtet werden. Will man solche linearen Kontraste im Rahmen der multiplen Regressionsanalyse untersuchen, so geht dies sehr einfach. Für jeden linearen Kontrast wird eine Codiervariable definiert. Die Gewichte der Mittelwerte im linearen Kontrast sind die

Tabelle 19.19 Bedeutung der Regressionskoeffizienten bei der Dummy- und Effektcodierung

Regressionskoeffizienten	Dummy-Codierung	Ungewichtete Effektcodierung	Gewichtete Effektcodierung
b_0	Mittelwert in der Kategorie, die auf allen Codiervariablen eine 0 aufweist (Referenzkategorie)	ungewichteter Mittelwert der Kategorienmittelwerte	gewichteter Mittelwert der Kategorienmittelwerte
b_j	Differenz der Mittelwerte der Kategorie j und der Referenzkategorie	Differenz des Mittelwerts der Kategorie j und des ungewichteten Gesamtmittelwerts	Differenz des Mittelwerts der Kategorie j und des gewichteten Gesamtmittelwerts

Werte der Codiervariablen. Diese Werte werden den einzelnen Bedingungen zugeordnet. Die multiple Regressionsanalyse mit diesen Kontrasten führt zum selben Ergebnis wie die Analyse der Kontraste in Abschnitt 13.1.13.

19.11.4 Inferenzstatistische Absicherung der Regressionsparameter

Für die inferenzstatistische Absicherung der Regressionsparameter bei der multiplen Regression mit kategorialen unabhängigen Variablen sind keine zusätzlichen Annahmen notwendig. Die Codiervariablen sind ganz »normale« Prädiktoren in einer multiplen Regressionsanalyse. In dem zugrunde liegenden Populationsmodell repräsentieren die Regressionsgewichte die entsprechenden Unterschiede zwischen den Populationsmittelwerten. Die Konfidenzintervalle, statistischen Tests und die Bestimmung der Stichprobengröße folgen dem in Abschnitt 19.7 beschriebenen Vorgehen.

Wir beschränken uns daher darauf, in Tabelle 19.20 die geschätzten Parameter, ihre Standardfehler und t -Werte anzugeben. In allen drei Fällen erhält man einen Determinationskoeffizienten von $R^2 = 0,009$, der auf einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ signifikant von 0 verschieden ist ($F = 11,015$; $df_1 = 2$, $df_2 = 2.381$; $p < 0,001$). Der Wert des Determinationskoeffizienten ist exakt gleich dem Wert von $\hat{\eta}^2$, den man erhalten würde, wenn die Daten mit einer einfaktoriellen Varianzanalyse (s. Abschn. 13.1.4) ausgewertet werden würden. Auch der F -Wert ist exakt identisch. Die Größe des Determinationskoeffizienten zeigt an, dass der Effekt sehr klein ist. Nur 0,9 % der Unterschiede in der materiellen Situation werden durch die Größe der Stadt, in der man wohnt, bestimmt.

Die Signifikanztests in Tabelle 19.20 zeigen in Bezug auf die Dummy-Codierung, dass die mittlere materielle Situation in der Kleinstadt (b_0) bedeutsam von 0 verschieden ist. Auch unterscheidet sich die materielle Situation von Personen, die in Großstädten wohnen, signifikant von solchen, die in Klein-

Tabelle 19.20 Ergebnisse der Parameterschätzungen für die Dummy- und die Effektcodierungen

a Dummy-Codierung				
	Wert	Standardfehler	t-Wert	p-Wert
b_0	-0,054	0,024	-2,224	0,026
b_1	0,132	0,033	4,007	< 0,001
b_2	0,001	0,034	0,042	0,967
b Ungewichtete Effektcodierung				
	Wert	Standardfehler	t-Wert	p-Wert
b_0	-0,010	0,014	-0,725	0,469
b_1	0,088	0,019	4,694	< 0,001
b_2	-0,043	0,019	-2,214	0,027
c Gewichtete Effektcodierung				
	Wert	Standardfehler	t-Wert	p-Wert
b_0	-0,004	0,014	-0,316	0,752
b_1	0,082	0,017	4,693	< 0,001
b_2	-0,049	0,020	-2,444	0,015

städten wohnen (b_1). Nicht bedeutsam hingegen sind die Unterschiede zwischen Mittel- und Kleinstädten hinsichtlich der mittleren materiellen Situation ihrer Einwohner (b_2).

Die Ergebnisse der Effektcodierungen unterscheiden sich kaum zwischen der gewichteten und der ungewichteten Form, da auch der Unterschied zwischen dem ungewichteten Mittelwert und dem gewichteten Mittelwert über die drei Kategorien hinweg nur relativ gering ist (s. Tab. 19.16). Die Ergebnisse zeigen zunächst, dass der Mittelwert der materiellen Situation nicht von 0 verschieden ist. Dies verwundert nicht, da die abhängige Variable ja auf der Basis z -transformierter Indikatoren der materiellen Situation definiert wurde und Standardwerte immer einen Mittelwert von 0 haben. Dass der Koeffizient b_0 nicht exakt gleich 0 ist, ist auf einzelne fehlende Werte zurückzuführen. Beide Regressionsgewichte b_1 und b_2 sind von 0 verschieden, was inhaltlich bedeutet, dass die materielle Situation von Großstadtbewohnern überdurchschnittlich und die von Mittelstadtbewohnern unterdurchschnittlich ist.

19.11.5 Analyse mehrerer kategorialer unabhängiger Variablen

Die Erweiterung auf mehr als eine kategoriale unabhängige Variable ist einfach. Zunächst muss man sich für eine Form der Codierung entscheiden. Dann müssen für alle unabhängigen Variablen Codiervariablen definiert werden. Um Interaktionen zwischen den unabhängigen Variablen zu überprüfen, müssen wie bei der moderierten Regressionsanalyse Produktvariablen in die Regressionsgleichung aufgenommen werden. Die Produktvariablen erhält man, indem man alle Codiervariablen, die zur Codierung der Bedingungen der ersten unabhängigen Variablen benötigt werden, mit allen Codiervariablen, die zur Codierung der Bedingungen der zweiten unabhängigen Variablen benötigt werden, multipliziert. Um die Äquivalenz zwischen Varianzanalyse und Regressionsanalyse mit kategorialen unabhängigen Variablen deutlich zu machen, bezeichnen wir die unabhängigen Variablen wie in der Varianzanalyse als Faktoren. Wir erklären das Vorgehen, indem wir das Beispiel im letzten Abschnitt um einen zweiten Faktor erweitern, der zwei Kategorien (Stufen) aufweist. Der zweite Faktor (Region) zeigt an, ob die untersuchte Person in den alten

Bundesländern (West) oder in den neuen Bundesländern (Ost) lebt. Wir zeigen also im Folgenden, wie man eine zweifaktorielle Varianzanalyse mit dem Faktor A (Größenstatus einer Stadt, dreistufig: klein, mittel, groß) und dem Faktor B (Region, zweistufig: Ost, West) und der abhängigen Variablen Y (materielle Situation) im Rahmen der multiplen Regressionsanalyse durchführen kann. Wir greifen auf die Dummy-Codierung zurück, könnten aber auch auf jede andere der beschriebenen Codierformen zurückgreifen. Das im Folgenden beschriebene Vorgehen ist direkt auf alle anderen Codierformen übertragbar. Zunächst müssen für den zweiten Faktor Codiervariablen eingeführt werden. Da dieser nur über zwei Kategorien verfügt, reicht eine Codiervariable aus, um die Information, die in dem Ost-West-Vergleich steckt, zu codieren. Wir weisen den neuen Bundesländern den Wert 0 und den alten Bundesländern den Wert 1 zu; damit ist die Region »neue Bundesländer« die Referenzkategorie. In Tabelle 19.21 sind die Werte zwölf ausgewählter Personen auf der abhängigen Variablen Y sowie auf den Codiervariablen angegeben.

Codiervariablen für die unabhängigen Merkmale

Die erste Spalte in Tabelle 19.21 enthält die Werte auf der abhängigen Variablen (materielle Situation). Spalte 2 und 3 enthalten die zugewiesenen Werte auf den beiden Dummy-Variablen zur Codierung der Stufen auf dem Faktor »Ortsgröße«, die wir schon kennengelernt haben. Wir fügen einen Index A hinzu, um deutlich zu machen, dass sich diese beiden Codiervariablen auf den ersten Faktor beziehen. Spalte 4 enthält die zugewiesenen Werte auf der Dummy-Variablen zur Codierung der Stufen auf dem Faktor »Region«. Sie erhält den Index B , um deutlich zu machen, dass sie den zweiten Faktor repräsentiert. Jeder Faktor wird durch eine Menge von Codiervariablen repräsentiert, wobei in jeder Menge die Anzahl der Codiervariablen gleich der Anzahl der Kategorien (Stufen) minus 1 ist. Würde man weitere kategoriale unabhängige Variablen berücksichtigen, würde man entsprechend weitere Mengen von Codiervariablen hinzufügen.

Codiervariablen für die Interaktion

Die Spalten 5 und 6 enthalten die Codiervariablen für die Interaktion zwischen den beiden Faktoren A und B . Die Anzahl der Codiervariablen, die für die Codierung der Interaktion notwendig sind, ergibt sich aus dem Produkt der Anzahl der Codiervariablen des einen Faktors mit der Anzahl der Codiervariablen des ande-

Tabelle 19.21 Multiple Regression mit zwei kategorialen unabhängigen Variablen und ihrer Interaktion; Dummy-Codierung

	Faktor A Ortsgröße		Faktor B Region	Interaktion			
Y	X _{A1}	X _{A2}	X _B	X _{A1 • B}	X _{A2 • B}		
1,98	1	0	1	1	0	Großstadt	West
−0,28	1	0	1	1	0		
−0,15	0	1	1	0	1	Mittelstadt	
0,55	0	1	1	0	1		
−0,30	0	0	1	0	0	Kleinstadt	
−0,18	0	0	1	0	0		
−0,38	1	0	0	0	0	Großstadt	Ost
−0,40	1	0	0	0	0		
−0,53	0	1	0	0	0	Mittelstadt	
0,16	0	1	0	0	0		
−0,48	0	0	0	0	0	Kleinstadt	
−0,15	0	0	0	0	0		

ren Faktors. In unserem Beispiel haben wir zwei Codierungsvariablen für den Faktor »Größenstatus« und eine Codierungsvariable für den Faktor »Region«, wir benötigen daher $2 \cdot 1 = 2$ Codierungsvariablen für die Interaktion. Die Werte auf diesen beiden Codierungsvariablen ergeben sich einfach aus dem Produkt der Werte auf denjenigen Codierungsvariablen, aus denen das Produkt gebildet wurde. Somit ergeben sich die Werte der Interaktionscodierungsvariablen $X_{A1 \cdot B}$ aus dem Produkt der dummy-codierten Variablen X_{A1} und der Codierungsvariablen X_B . Die Werte der zweiten Codierungsvariablen der Interaktion ergeben sich aus dem Produkt der Werte der Codierungsvariablen X_{A2} und der Codierungsvariablen X_B (s. Tab. 19.21).

Bedeutung der Regressionskoeffizienten

Bezogen auf unser Beispiel lautet die Gleichung der Regressionsanalyse:

$$\hat{Y} = b_0 + b_1 \cdot X_{A1} + b_2 \cdot X_{A2} + b_3 \cdot X_B + b_4 \cdot X_{A1 \cdot B} + b_5 \cdot X_{A2 \cdot B} \quad (\text{F 19.77})$$

Berechnet man eine multiple Regressionsanalyse auf der Grundlage der Daten aus der Untersuchung von Schmitt und Maes (2001) mit $n = 1.942$ Personen und legt die in Tabelle 19.21 dargestellten Codierungen zugrunde, erhält man die folgenden geschätzten Regressionskoeffizienten:

$$\begin{aligned} \hat{Y} = & -0,192^{**} + 0,054 \cdot X_{A1} + 0,049 \cdot X_{A2} \\ & (0,032) \quad (0,045) \quad (0,045) \\ & + 0,352^{**} \cdot X_B + 0,033 \cdot X_{A1 \cdot B} - 0,182^* \cdot X_{A2 \cdot B} \\ & (0,052) \quad (0,069) \quad (0,072) \end{aligned} \quad (\text{F 19.78})$$

Die Zahlen (in Klammern) unter den geschätzten Parametern geben den Standardfehler des jeweiligen Regressionskoeffizienten an. Die Sternchen repräsentieren das Ergebnis des dazugehörigen t -Tests auf Absicherung des jeweiligen Regressionskoeffizienten gegen 0. Ein Sternchen zeigt an, dass der p -Wert kleiner oder gleich 0,05 ist, zwei Sternchen spiegeln ein $p \leq 0,01$ wider. Fehlt ein Sternchen, zeigt dies an, dass der Wert auf einem Signifikanzniveau von $\alpha = 5\%$ (zweiseitig) nicht signifikant von 0 verschieden ist.

Die Regressionskoeffizienten der Dummy-Variablen zur Codierung des Faktors »Größenstatus« unterscheiden sich von den Koeffizienten, die wir für das Modell ohne den Ost-West-Vergleich erhalten haben (s. Formel F 19.76). Dies ist darauf zurückzuführen, dass sich die Bedeutung der Regressionsgewichte der Dummy-Variablen zur Codierung des Faktors »Größenstatus« ändert, wenn weitere Prädiktoren hinzukommen. Die Bedeutung der Regressionsanalyse ergibt sich aus dem Modell

der multiplen Regressionsanalyse, so wie wir es auch für die metrischen unabhängigen Variablen eingeführt haben. Neben diesem Ansatz gibt es im Falle korrelierter

Faktoren (sog. nonorthogonale Varianzanalyse, s. Abschn. 13.2.9) noch weitere Ansätze, die wir nicht weiter behandeln werden (s. folgenden Kasten).

Vertiefung

Nonorthogonale Varianzanalyse: Vier Typen von Quadratsummen zur Überprüfung spezifischer Effekte

Bei der Analyse kategorialer unabhängiger Variablen gibt es neben der Zerlegung der Quadratsummen, auf die wir im Rahmen der multiplen Regressionsanalyse in den Abschnitten 19.6 und 19.7 zurückgegriffen haben (sog. Quadratsumme vom Typ III), drei weitere Typen von Quadratsummen, die sich nur dann nicht unterscheiden, wenn die unabhängigen Variablen unabhängig voneinander sind (sog. orthogonales Design). Liegen Abhängigkeiten zwischen den unabhängigen Variablen vor (nonorthogonales Design), so kann in unterschiedlicher Weise mit diesen Abhängigkeiten umgegangen werden. Hierzu gibt es eine umfangreiche Literatur (Stichwort: nonorthogonale Varianzanalyse), die wir in einem einführenden Lehrbuch nicht behandeln können (s. hierzu ausführlich Steyer, 1979; Werner, 1997). Wir wollen sie jedoch kurz skizzieren und anhand unseres Beispiels zeigen, was sich dahinter verbirgt. Die vier Quadratsummen unterscheiden sich darin, wie die Abhängigkeiten zwischen den unabhängigen Merkmalen (Faktoren) korrigiert werden. Diese vier Typen von Quadratsummen werden von Statistikprogrammen unter dem Allgemeinen Linearen Modell berechnet, z. B. von dem Computerprogramm SPSS unter der Option »Modell«. Die Voreinstellung ist typischerweise »Typ III«, da diese Quadratsumme dem üblichen regressionsanalytischen Modell entspricht. Bei jedem Typ werden für die Regressionskonstante, jeden Faktor sowie für jede Interaktion zwischen Faktoren eine Quadratsumme, eine mittlere Quadratsumme sowie ein F -Wert und ein p -Wert bestimmt. Wir lassen im Folgenden die Quadratsumme der Regressionskonstanten außen vor und widmen uns den unabhängigen Merkmalen (Faktoren).

Typ I. Bei der Quadratsumme vom Typ I hängen die Quadratsummen von der Reihenfolge der Aufnahme der unabhängigen Merkmale (Faktoren) ins Modell ab. Diese Reihenfolge muss bei Statistikprogrammen spezifiziert werden (z. B. bei SPSS unter der Option »Modell«). Die erste Quadratsumme repräsentiert

den Einfluss des ersten Faktors, der in das Modell aufgenommen wurde. Der Einfluss des Faktors wird nicht korrigiert, weder in Bezug auf seine Abhängigkeit mit den anderen Faktoren noch mit den Interaktionsvariablen. Sein Einfluss repräsentiert also auch die Variationsquellen, die er mit den anderen unabhängigen Variablen teilt. Die Quadratsumme des zweiten Faktors, der ins Modell aufgenommen wurde, kennzeichnet den Einfluss des zweiten Faktors, der um die Abhängigkeit mit dem ersten Faktor korrigiert wurde, nicht aber in Bezug auf seine Abhängigkeit mit den Interaktionsvariablen. Die Quadratsumme zeigt also an, was der zweite Faktor über den ersten Faktor hinaus erklärt. Hat man mehr als zwei Faktoren, werden die weiteren Faktoren in derselben Weise behandelt (korrigiert). Schließlich repräsentiert die Quadratsumme der Interaktion den Anteil der Variation, der durch die Interaktion bedingt ist, nachdem die Haupteffekte auspartialisiert wurden. In unserem Beispiel gibt es zwei mögliche Reihenfolgen, mit denen die Faktoren aufgenommen werden können: (1) zuerst die Stadtgröße und dann die Region oder (2) zuerst die Region und dann die Stadtgröße. Für welche Reihenfolge man sich entscheidet, muss theoretisch gut begründet sein. Angenommen, in unserem Beispiel wären die Ortsgröße und die Region stark voneinander abhängig, da es in den alten Bundesländern mehr Groß- und in den neuen Bundesländern mehr Kleinstädte gäbe, und man fände, dass es sowohl einen West-Ost- als auch einen Großstadt-Kleinstadt-Unterschied in der materiellen Situation gebe, dann stellte sich die Frage, ob der Unterschied in der materiellen Situation eher auf den West-Ost-Unterschied oder den Großstadt-Kleinstadt-Unterschied zurückführbar wäre. Würde man argumentieren, dass der West-Ost-Unterschied der zentrale (»kausale«) Wirkfaktor sei und der Großstadt-Kleinstadt-Unterschied nur dadurch zustande komme, dass es im Westen mehr Großstädte gäbe, würde man den Faktor Region als Ersten aufnehmen, dann den Faktor Ortsgröße, um zu sehen, ob diese noch über den Regionsunterschied hinaus etwas

erklärt, und schließlich die Interaktion. Die Quadratsumme vom Typ I ist dann geeignet, wenn die Abhängigkeit der Variablen substanzielle Unterschiede in den Zellenbesetzungen widerspiegelt, die nicht einfach auf unsystematischen Ausfall zurückführbar sind und es klare Hypothesen über eine Kausalkette gibt, die getestet werden sollen. Das Vorgehen bei dieser Quadratsumme entspricht dem Vorgehen der multiplen Regressionsanalyse, wenn blockweise unabhängige Variablen aufgenommen werden. Die Quadratsumme vom Typ I ist für experimentelle Designs mit Zufallsausfällen nicht geeignet, da es hier keine klare hierarchische Ordnung gibt.

Typ II. Bei der Quadratsumme vom Typ II werden die Haupteffekte in Bezug auf ihre Abhängigkeit von den anderen Haupteffekten korrigiert, nicht aber in Bezug auf ihre Abhängigkeit von den Interaktionsvariablen. Die Quadratsumme einer der Haupteffekte spiegelt also wider, was dieser Faktor über die anderen Faktoren hinaus erklärt. Es findet also eine wechselseitige Korrektur statt, wie es bei einer multiplen Regressionsanalyse der Fall ist, wenn keine Interaktionsvariablen aufgenommen wurden. Die Quadratsumme der Interaktion ist in Bezug auf alle Haupteffekte korrigiert. Diese Quadratsumme ist dann geeignet, wenn es keine klare kausale Ordnung gibt und keine Interaktion vorliegt.

Typ III. Die Quadratsumme vom Typ III entspricht der üblichen multiplen Regressionsanalyse, wenn alle unabhängigen Variablen aufgenommen wurden. Alle unabhängigen Variablen werden in Bezug auf ihre Abhängigkeit von allen anderen unabhängigen Variablen korrigiert. Die Quadratsumme eines Fak-

tors spiegelt wider, welchen Beitrag dieser Faktor zur Varianzaufklärung über die anderen Faktoren und die Interaktionen hinaus leistet. Die Quadratsumme von diesem Typ ist die Voreinstellung bei vielen Computerprogrammen. Sie ist insbesondere für experimentelle Studien geeignet, bei denen die Abhängigkeiten durch zufällige Ausfälle entstehen. Wir greifen in unserem Anwendungsfall auf diesen Quadratsummentyp zurück, da die Studie so angelegt war, dass die beiden Faktoren voneinander unabhängig sein sollten, indem für jede Kombination (Zelle) der beiden Faktoren gleich große Zufallsstichproben gezogen wurden. Die Nonorthogonalität entsteht dadurch, dass die Rücklaufquoten von den Personen, die kontaktiert wurden, nicht für alle Zellen gleich waren.

Typ IV. Dieser Quadratsummentyp ist dann geeignet, wenn bestimmte Zellkombinationen bewusst nicht besetzt sind (sog. strukturelle Nullhäufigkeiten). Sind – wie in unserem Beispiel – alle Zellen besetzt, entspricht diese Quadratsumme der Quadratsumme vom Typ III. Wir behandeln den Fall struktureller Nullhäufigkeiten nicht. Hierzu sei auf Werner (1997) verwiesen.

Kausale Effekte. Die einzelnen Quadratsummen sind mit spezifischen Interpretationen in Bezug auf die Bedeutung von Haupteffekten verknüpft, die ausführlich bei Werner (1997) behandelt werden. Sie erlauben jedoch nicht die Überprüfung aller Kausalhypothesen, die theoretisch sinnvoll sind. Will man anhand nicht-experimenteller Studien Kausalhypothesen untersuchen, sei die Lektüre der Arbeiten von Steyer (1992, 2003) empfohlen.

Wir werden im Folgenden die Bedeutung jedes einzelnen Regressionskoeffizienten herleiten und das Ergebnis der empirischen Untersuchung interpretieren. Die Bedeutung der Regressionskoeffizienten kann man sich leicht veranschaulichen, wenn man Werte für die verschiedenen X-Variablen in Gleichung F 19.77 einsetzt.

Bedeutung von b_0 . Der Achsenabschnitt entspricht dem erwarteten \hat{y} -Wert, wenn alle Codiervariablen den Wert 0 annehmen. Dies ist der Mittelwert der materiellen Situation aller Personen, die in einer Kleinstadt in den

neuen Bundesländern leben. Der Wert von $b_0 = -0,192$ ist signifikant von 0 verschieden. Da ein z-Wert von 0 dem Gesamtmittelwert entspricht, ist die materielle Situation in ostdeutschen Kleinstädten unterdurchschnittlich und liegt um 0,192 Standardabweichungen unter dem Mittelwert.

Bedeutung von b_1 . Das Regressionsgewicht b_1 entspricht der Differenz aus dem Mittelwert der materiellen Situation in einer ostdeutschen Großstadt und einer ostdeutschen Kleinstadt. Dies folgt unmittelbar aus dem

Codierschema in Tabelle 19.21. Personen, die in einer ostdeutschen Großstadt leben, haben auf der ersten Codiervariablen einen Wert von 1, auf allen anderen einen Wert von 0. Für alle Personen dieser Gruppe setzt sich ihr erwarteter \hat{y} -Wert (also der Gruppenmittelwert) additiv zusammen aus dem Achsenabschnitt b_0 und dem Regressionsgewicht b_1 : $\hat{y}_{\text{Großstadt-Ost}} = b_0 + b_1$. Hieraus folgt, dass das Regressionsgewicht die Differenz aus dem Mittelwert in einer ostdeutschen Großstadt und einer ostdeutschen Kleinstadt darstellt. Dieser Wert beträgt in unserem Beispiel $b_1 = 0,054$ und ist nicht bedeutsam von 0 verschieden. In den neuen Bundesländern unterscheiden sich Klein- und Großstädte somit nicht in ihrer mittleren materiellen Situation.

Bedeutung von b_2 . Das Regressionsgewicht b_2 entspricht der Differenz aus dem Mittelwert der materiellen Situation in einer ostdeutschen Mittelstadt und einer ostdeutschen Kleinstadt. Dies folgt ebenfalls aus dem Codierschema in Tabelle 19.21. Personen, die in einer ostdeutschen Mittelstadt leben, haben auf der zweiten Codiervariablen einen Wert von 1, auf allen anderen einen Wert von 0. Daher erhält man für den erwarteten \hat{y} -Wert in ostdeutschen Mittelstädten: $\hat{y}_{\text{Mittelstadt-Ost}} = b_0 + b_2$. Für unser Beispiel ergibt sich ein Wert von $b_2 = 0,049$, der ebenfalls nicht signifikant von 0 verschieden ist. Auch die mittlere materielle Situation ostdeutscher Mittelstädte unterscheidet sich nicht von derjenigen ostdeutscher Kleinstädte.

Bedeutung von b_3 . Das Regressionsgewicht b_3 entspricht der Differenz zwischen der mittleren materiellen Situation in westdeutschen Kleinstädten und der mittleren materiellen Situation in ostdeutschen Kleinstädten. Die mittlere materielle Situation von Personen in westdeutschen Kleinstädten ergibt sich auf der Basis des Codierschemas in Tabelle 19.21 nach Gleichung F 19.77 zu $\hat{y}_{\text{Kleinstadt-West}} = b_0 + b_3$. In unserem Beispiel erhalten wir den Wert $b_3 = 0,352$; dieser Wert ist signifikant von 0 verschieden. Die mittlere materielle Situation ist in westdeutschen Kleinstädten also signifikant höher als in ostdeutschen Kleinstädten.

Bedeutung von b_4 . Um die Bedeutung des Regressionsgewichts b_4 herzuleiten, setzen wir die Werte von Personen in westdeutschen Großstädten (s. Tab. 19.21) in die Gleichung F 19.77 ein und erhalten:

$$\hat{y}_{\text{Großstadt-West}} = b_0 + b_1 + b_3 + b_4$$

Durch Umformen und Einsetzen der Bedeutung von b_0 , b_1 und b_3 erhält man:

$$\begin{aligned} b_4 &= \hat{y}_{\text{Großstadt-West}} - b_0 - b_1 - b_3 \\ &= \hat{y}_{\text{Großstadt-West}} - \hat{y}_{\text{Kleinstadt-Ost}} - \hat{y}_{\text{Großstadt-Ost}} \\ &\quad + \hat{y}_{\text{Kleinstadt-Ost}} - \hat{y}_{\text{Kleinstadt-West}} + \hat{y}_{\text{Kleinstadt-Ost}} \\ &= \hat{y}_{\text{Großstadt-West}} - \hat{y}_{\text{Großstadt-Ost}} - \hat{y}_{\text{Kleinstadt-West}} \\ &\quad + \hat{y}_{\text{Kleinstadt-Ost}} \\ &= (\hat{y}_{\text{Großstadt-West}} - \hat{y}_{\text{Großstadt-Ost}}) \\ &\quad - (\hat{y}_{\text{Kleinstadt-West}} - \hat{y}_{\text{Kleinstadt-Ost}}) \end{aligned}$$

Das Regressionsgewicht kontrastiert somit Unterschiede zwischen west- und ostdeutschen Großstädten mit Unterschieden in west- und ostdeutschen Kleinstädten. Der Wert des Regressionsgewichts ist gleich 0, wenn die Differenz in der mittleren materiellen Situation zwischen west- und ostdeutschen Städten nicht davon abhängt, ob es sich um eine Groß- oder Kleinstadt handelt (oder alternativ: wenn die Differenz in der mittleren materiellen Situation zwischen Klein- und Großstädten nicht von der Region abhängt). Ist hingegen der Wert des Regressionskoeffizienten von 0 verschieden, zeigt dies an, dass der West-Ost-Effekt von der Größe der Stadt abhängt (moderiert wird bzw. alternativ: dass der Unterschied zwischen Klein- und Großstädten von der Region moderiert wird). Hierdurch kommt der Interaktionseffekt zum Ausdruck: Der Effekt eines Faktors auf die abhängige Variable hängt von der Ausprägung auf dem anderen Faktor ab. In unserer Untersuchung ergibt sich ein Wert von $b_4 = 0,033$, der auf einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ nicht signifikant von 0 verschieden ist. Der Unterschied zwischen west- und ostdeutschen Städten in der mittleren materiellen Situation hängt nicht davon ab, ob es sich um Groß- oder Kleinstädte handelt. (Oder alternativ: Der Unterschied zwischen Klein- und Großstädten ist zwischen den alten und den neuen Bundesländern nicht signifikant.)

Bedeutung von b_5 . Die Bedeutung des Regressionsgewichts b_5 leiten wir in gleicher Weise her wie die Bedeutung des Regressionsgewichts b_4 . Setzen wir die Werte von Personen in westdeutschen Mittelstädten (s. Tab. 19.21) in die Gleichung F 19.77 ein, erhalten wir:

$$\hat{y}_{\text{Mittelstadt-West}} = b_0 + b_2 + b_3 + b_5$$

Durch Umformen und Einsetzen der Bedeutung von b_0 , b_2 und b_3 ergibt sich:

$$\begin{aligned}
 b_5 &= \hat{y}_{\text{Mittelstadt-West}} - b_0 - b_2 - b_3 \\
 &= \hat{y}_{\text{Mittelstadt-West}} - \hat{y}_{\text{Kleinstadt-Ost}} - \hat{y}_{\text{Mittelstadt-Ost}} \\
 &\quad + \hat{y}_{\text{Kleinstadt-Ost}} - \hat{y}_{\text{Kleinstadt-West}} + \hat{y}_{\text{Kleinstadt-Ost}} \\
 &= \hat{y}_{\text{Mittelstadt-West}} - \hat{y}_{\text{Mittelstadt-Ost}} - \hat{y}_{\text{Kleinstadt-West}} \\
 &\quad + \hat{y}_{\text{Kleinstadt-Ost}} \\
 &= (\hat{y}_{\text{Mittelstadt-West}} - \hat{y}_{\text{Mittelstadt-Ost}}) \\
 &\quad - (\hat{y}_{\text{Kleinstadt-West}} - \hat{y}_{\text{Kleinstadt-Ost}})
 \end{aligned}$$

Das Regressionsgewicht b_5 vergleicht Unterschiede zwischen west- und ostdeutschen Mittelstädten mit Unterschieden zwischen west- und ostdeutschen Kleinstädten (oder alternativ: zwischen Klein- und Mittelstädten in den neuen Bundesländern mit Unterschieden zwischen Klein- und Mittelstädten in den alten Bundesländern). Ist das Regressionsgewicht von 0 verschieden, so bedeutet das, dass der West-Ost-Unterschied in Kleinstädten anders ausfällt als in Mittelstädten (oder alternativ: dass der Unterschied zwischen Klein- und Mittelstädten in den neuen Bundesländern anders ausfällt als in den alten Bundesländern). Der Wert von $b_5 = -0,182$ in unserer Untersuchung ist negativ und in der Tat signifikant von 0 verschieden (bei $\alpha = 5\%$). Er zeigt an, dass sich der West-Ost-Unterschied in der materiellen Situation in Kleinstädten deutlicher zeigt als in Mittelstädten.

Bedeutung der Regressionskoeffizienten im restringierten Modell

Wir haben in den vorherigen Abschnitten gezeigt, dass sich die Regressionskoeffizienten als Funktionen der empirischen Mittelwerte ausdrücken lassen. Dies ist jedoch nur dann der Fall, wenn alle möglichen Effekte – also die Haupteffekte und die Interaktionseffekte – in der Regressionsgleichung enthalten sind. Ein solches Modell, das alle möglichen Effekte enthält, nennt man saturiertes Modell. Würde man z. B. die Interaktionseffekte aus dem Regressionsmodell entfernen, so lie-

ßen sich die Regressionskoeffizienten nur als Funktion der erwarteten \hat{y} -Werte formulieren. Die erwarteten \hat{y} -Werte können dann von den empirischen Mittelwerten abweichen. Nimmt man mögliche Effekte aus dem Modell heraus, so spricht man von einem restringierten Modell. Ist das restringierte Modell in der Population korrekt, so sind die Regressionsparameter in der Population als Funktionen der Populationsmittelwerte darstellbar. Selbst wenn das Modell in der Population gilt, werden sich aber die in der Stichprobe berechneten erwarteten \hat{y} -Werte aufgrund des Stichprobenfehlers von den empirischen Mittelwerten unterscheiden.

Inferenzstatistische Absicherung des Interaktionseffekts

In unserem Beispiel haben wir die einzelnen Regressionskoeffizienten auf Signifikanz geprüft. Im Rahmen der multiplen Regressionsanalyse kann man auch die Nullhypothese, dass weder die eine noch die andere Interaktionsvariable einen bedeutsamen Einfluss hat, überprüfen. In unserem Beispiel lautet diese Nullhypothese $H_0: \beta_4 = \beta_5 = 0$. Man überprüft sie, indem man zunächst in einem ersten Block die drei Codiervariablen der Faktoren A und B in die Regressionsgleichung aufnimmt und dann in einem zweiten Block die beiden Codiervariablen, die die Interaktion repräsentieren, hinzunimmt. Ist die Zunahme des Determinationskoeffizienten statistisch bedeutsam, muss die Nullhypothese, dass keine Interaktion vorliegt, verworfen werden. Die Ergebnisse dieser Analysen sind für unser Beispiel in Tabelle 19.22 zusammengestellt. Die Hinzunahme der beiden Interaktionscodiervariablen führt zu einer signifikanten Erhöhung des Determinationskoeffizienten. Durch die Interaktionsvariablen werden zusätzlich 0,6 % an Varianz erklärt.

Vorgehen bei mehr als zwei unabhängigen Merkmalen

Hat man mehr als zwei unabhängige Merkmale, geht man in derselben Weise wie beschrieben vor. Man codiert zunächst jedes unabhängige Merkmal. Um die

Tabelle 19.22 Regressionsanalyse mit Dummy-Variablen: Überprüfung der Interaktionshypothese

Unabhängige Variablen	R	R ²	Ünderungsstatistiken				
			Ünderung in R ²	F-Wert	df ₁	df ₂	p-Wert
Ortsgröße, Ost – West	0,253	0,064	0,064	44,316	3	1.938	< 0,001
Ortsgröße, Ost – West, Interaktion	0,264	0,070	0,006	5,578	2	1.936	0,004

Interaktion zu überprüfen, müssen dann alle Codiervariablen eines Merkmals mit allen Codiervariablen eines anderen Merkmals multipliziert werden. Um Interaktionen höherer Ordnung zu überprüfen, müssen die Codiervariablen von mehreren Merkmalen miteinander multipliziert werden etc.

19.11.6 Ordinale unabhängige Variablen

Wir haben bisher gezeigt, wie (1) metrische und (2) nominalskalierte kategoriale unabhängige Variablen in einer multiplen Regressionsanalyse behandelt werden können. Nominalskalierte unabhängige Variablen weisen keine Ordnung der Kategorien auf. In der Psychologie kommt aber häufig der Fall vor, dass man kategoriale Variablen vorliegen hat, bei denen die Kategorien geordnet sind. Ein Beispiel dafür sind Ratingskalen (s. hierzu z. B. Eid & Schmidt, 2014). Wie geht man mit solchen Variablen um, wenn man sie in die multiple Regressionsanalyse aufnehmen will? Typischerweise greift man zu einer von zwei Vorgehensweisen: Entweder man behandelt die Variablen als nominalskaliert und nimmt entsprechende Codiervariablen in die Regressionsanalyse auf. Dies ist gerechtfertigt, da ordinalskalierte Variablen immer auch nominalskalierte Variablen sind (s. Abschn. 5.1.1). Hat eine Ratingskala jedoch viele Kategorien, erhält man sehr viele Codiervariablen und zu schätzende Parameter. Die andere Möglichkeit besteht darin, sie als metrische Variablen zu behandeln. Dies setzt allerdings voraus, dass die ordinalen Variablen den Anforderungen einer metrischen Variablen genügen, d. h., dass die Abstände der Werte sinnvoll interpretierbar sind (s. Abschn. 5.1.1). Neben diesen beiden traditionellen Methoden der Behandlung ordinaler Variablen gibt es moderne statistische Verfahren, die es erlauben, das ordinale Skalenniveau der unabhängigen Variablen adäquat zu berücksichtigen (Tutz & Gertheiss, 2014).

19.12 Gemeinsame Analyse kategorialer und metrischer unabhängiger Variablen

In der multiplen Regressionsanalyse ist es nicht nur möglich, metrische und kategoriale unabhängige Va-

riablen getrennt zu analysieren, es gibt darüber hinaus auch die Möglichkeit, metrische Variablen und kategoriale Variablen gemeinsam in eine Analyse zu integrieren. Dies möchten wir anhand eines Beispiels aus der kulturvergleichenden Forschung illustrieren.

Beispiel

Positive Emotionen und Lebenszufriedenheit in China und den USA

In einer Untersuchung zur Lebenszufriedenheit wurde untersucht, inwieweit das Lebenszufriedenheitsurteil einer Person von der Anzahl der positiven Emotionen, die eine Person in einem gewissen Zeitraum erlebt hat, abhängt. Aus dieser Studie haben wir zu Illustrationszwecken zwei Nationen (China, USA) ausgewählt. Die Lebenszufriedenheit wurde mit der Satisfaction-with-Life-Skala (Pavot & Diener, 1993) erfasst, einer Skala, die aus fünf Items besteht, die auf einer siebenstufigen Antwortskala eingeschätzt werden. Der Summenwert dieser fünf Items ist der Wert der abhängigen Variablen in dieser Analyse. Der Wertebereich reicht von 5 bis 35; je höher der Wert, desto höher die Lebenszufriedenheit. Von Interesse ist die Frage, inwieweit die Lebenszufriedenheit von der Anzahl positiver Affekte (*PA*), die eine Person in einem bestimmten Zeitraum erlebt hat, sowie der Zugehörigkeit zu einer Nation abhängt. Der Wertebereich der Skala *PA* reicht von 0 (keine positiven Affekte) bis 24 (sehr häufig positive Affekte).

Wir werden zwei Modelle analysieren, zunächst ein Modell mit einer additiven Verknüpfung der beiden unabhängigen Variablen, und danach ein Modell, in dem zusätzlich die Interaktion zwischen der Anzahl positiver Affekte und der Nation berücksichtigt wird.

19.12.1 Additive Verknüpfung kategorialer und kontinuierlicher Variablen: Kovarianzanalyse

Verknüpft man die unabhängigen kategorialen und kontinuierlichen Variablen additiv, erhält man ein multiples Regressionsmodell, das äquivalent zu der sog. *Kovarianzanalyse* ist. In der Kovarianzanalyse werden keine Interaktionen zwischen den kontinuierlichen

und den kategorialen Variablen zugelassen, während Interaktionen zwischen den kategorialen Variablen erlaubt sind. Die Kovarianzanalyse ist eine Erweiterung der Varianzanalyse um kontinuierliche unabhängige Variablen, die *Kovariaten* genannt werden. Die Kovarianzanalyse wird v.a. aus zweierlei Gründen eingesetzt: Zum einen kann man untersuchen, ob eine oder mehrere kontinuierliche Variablen einen Einfluss auf die abhängige Variable haben, der über die Effekte der kategorialen Variablen hinausgeht. Zum anderen wird durch die kontinuierlichen Variablen zusätzliche Varianz erklärt, wodurch die Residualvarianz in der abhängigen Variablen verringert wird. Hierdurch verringert sich der Zähler in der Formel zur Berechnung des Standardfehlers des Regressionsgewichts einer kategorialen Variablen (s. Formel F 19.39). Die Hinzunahme weiterer unabhängiger Variablen verringert zwar auch den Nenner des Standardfehlers, da sich die Freiheitsgrade erhöhen und die Kovariaten möglicherweise mit der kategorialen unabhängigen Variablen korreliert sind. Ist die Verkleinerung des Zählers jedoch größer als die Verkleinerung des Nenners, so verringert sich durch die Hinzunahme der Kovariaten der Standardfehler des Regressionsgewichts einer kategorialen unabhängigen Variablen. Hierdurch erhöht sich der entsprechende *t*-Wert zur Überprüfung des Effekts einer kategorialen Variablen (s. Formel F 19.40). Durch die Hinzunahme von Kovariaten, die einen bedeutsamen Einfluss haben, kann sich somit die Teststärke (Power) zur Absicherung eines Effekts der kategorialen Variablen erhöhen. Aufgrund des Gewinns an Teststärke werden weniger Versuchspersonen benötigt, als wenn man den Effekt einer kategorialen Variablen ohne Kovariaten absichern würde. Dies macht man sich insbesondere in der experimentellen Forschung zunutze, indem man Kovariaten auswählt, die mit der abhängigen Variablen hoch korreliert sind und aufgrund der Randomisierung nicht systematisch mit den kategorialen unabhängigen Variablen (experimentellen Bedingungen) zusammenhängen. Wir behandeln die Kovarianzanalyse im Rahmen der multiplen Regressionsanalyse, da dies den Vorteil hat, die Annahme der Kovarianzanalyse, dass keine Interaktionen zwischen den kontinuierlichen und kategorialen Variablen vorliegt, zu überprüfen. Das Vorgehen bei der Kovarianzanalyse folgt einer einfachen Struktur.

Definition

Die **Kovarianzanalyse** ist in folgenden Schritten vorzunehmen:

- (1) Repräsentiere die kategorialen unabhängigen Variablen durch Codiervariablen.
- (2) Nimm die Codiervariablen und die kontinuierlichen unabhängigen Variablen in die multiple Regressionsanalyse auf.
- (3) Lasse keine Interaktionen zwischen kategorialen und kontinuierlichen Variablen zu.

In unserem Beispiel codieren wir zunächst die Zugehörigkeit zu einer Nation. Wir haben hierzu die Dummy-Codierung gewählt. Personen, die in China leben, wird der Wert 0, Personen, die in den USA leben, der Wert 1 zugeordnet. Sowohl die metrische Variable positiver Affekt ($PA: X_1$) als auch die kategoriale Variable Nation (X_2) können nun als zwei unabhängige Variablen in die Analyse aufgenommen werden. Das multiple Regressionsmodell mit der abhängigen Variablen Lebenszufriedenheit ($LZ: Y$) und den beiden unabhängigen Variablen wird durch folgende Regressionsgleichung beschrieben:

$$\widehat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$$

bzw.

$$\widehat{LZ} = b_0 + b_1 \cdot PA + b_2 \cdot Nation$$

Eine Regressionsanalyse ($n = 998$) ergab folgende Regressionsgleichung:

$$\widehat{LZ} = 11,612^{**} + 0,699^{**} \cdot PA + 2,971^{**} \cdot Nation$$

(0,395) (0,046) (0,449)

Die Zahlen (in Klammern) unter den geschätzten Parametern geben den Standardfehler des jeweiligen Regressionskoeffizienten an. Teilt man den geschätzten Wert durch den Standardfehler, erhält man wie bei jeder anderen Regressionsanalyse den Wert einer *t*-verteilten Prüfgröße. Die Sternchen repräsentieren den *p*-Wert des dazugehörigen *t*-Tests. Ein Sternchen bedeutet wiederum, dass der *p*-Wert kleiner oder gleich 0,05 ist, zwei Sternchen zeigen ein $p \leq 0,01$ an. Alle drei Regressionskoeffizienten sind signifikant von 0 verschieden, legt man a priori ein $\alpha = 0,05$ fest. Wie lassen sich nun die Regressionskoeffizienten interpretieren?

Bedeutung von b_0 . Der Achsenabschnitt zeigt den erwarteten \hat{y} -Wert für Personen in China ($X_2 = 0$) an, die einen Wert von 0 auf der Variablen PA haben. Der Wert $X_1 = 0$ besagt, dass keine positiven Affekte erlebt werden. Für diese Personen erwarten wir einen Lebenszufriedenheitswert von $\hat{y} = 11,612$.

Bedeutung von b_1 . Das Regressionsgewicht von $b_1 = 0,699$ zeigt an, dass sowohl in China als auch in den USA die Lebenszufriedenheit ansteigt, wenn mehr positive Affekte erlebt werden. Mit einer Zunahme der positiven Affekte um den Wert 1 erwartet man in beiden Nationen einen mittleren Zuwachs der Lebenszufriedenheit von 0,699.

Bedeutung von b_2 . Das Regressionsgewicht von $b_2 = 2,971$ zeigt an, dass Personen, die gleich viel positive Affekte erleben, in den USA einen um den Wert 2,971 höheren erwarteten Lebenszufriedenheitswert aufweisen als in China. Es ist der bedingte Effekt der nationalen Zugehörigkeit, der mittlere Unterschied zwischen den beiden Nationen in der Lebenszufriedenheit, der nicht darauf zurückgeführt werden kann, dass sich beide Nationen auch im Erleben positiver Affekte unterscheiden.

Adjustierte Mittelwerte

In unserem Beispiel unterscheiden sich die beiden Nationen sowohl in ihren Mittelwerten für die Lebenszufriedenheit als auch in ihren Mittelwerten für die Häufigkeit positiver Affekte. Wir berichten in Tabelle 19.23 die Mittelwerte in beiden Nationen.

Tabelle 19.23 Mittelwerte für die Lebenszufriedenheit und die Häufigkeit positiver Affekte in China und den USA

	\bar{x}_1 (positiver Affekt)	\bar{y} (Lebenszufriedenheit)
China	6,941	16,464
USA	13,000	23,670
Gesamt	9,624	19,655

Man sieht, dass US-Amerikaner sowohl in der Lebenszufriedenheit als auch in der Häufigkeit positiver Affekte höhere Mittelwerte aufweisen als Chinesen. Um beide Nationen in Bezug auf ihre mittlere Lebenszufriedenheit zu vergleichen, insoweit sie nicht auf das

Erleben positiver Affekte zurückgeführt werden kann, berechnet man die adjustierten Mittelwerte.

Definition

Ein **adjustierter Mittelwert** ist der Mittelwert der abhängigen Variablen einer bestimmten Gruppe (Kategorie einer kategorialen unabhängigen Variablen), der anhand der Regressionsgleichung in dieser Gruppe für den Gesamtmittelwert der Kovariaten bestimmt wird.

Adjustiert bedeutet, dass wir Personen gleicher Ausprägung auf den Kovariaten betrachten und dadurch Unterschiede zwischen Personen konstant halten. In unserem Beispiel berechnen wir für beide Nationen die adjustierten Mittelwerte der Lebenszufriedenheit für den Gesamtmittelwert der Kovariaten positiver Affekte $\bar{x}_1 = 9,624$:

$$\hat{y}_{\text{China}} = 11,612 + 0,699 \cdot 9,624 = 18,339$$

$$\hat{y}_{\text{USA}} = 11,612 + 0,699 \cdot 9,624 + 2,971 = 21,310$$

Man sieht, dass der Mittelwertsunterschied in der Lebenszufriedenheit zwischen den USA und China kleiner geworden ist, wenn man Personen vergleicht, die gleich häufig positive Affekte erleben. Während der Mittelwertsunterschied in der Lebenszufriedenheit ohne Berücksichtigung der Kovariaten $23,670 - 16,464 = 7,206$ beträgt, ergibt sich für die adjustierten Mittelwerte ein Unterschied von 2,971. Die Differenz der adjustierten Mittelwerte entspricht genau dem Regressionsgewicht der Variablen Nation. Dies ist nicht erstaunlich, da dieses Regressionsgewicht den Unterschied zwischen den beiden Regressionsgeraden angibt, d.h. den Wert, um den die Regressionsgerade in den USA über der Regressionsgeraden in China liegt (s. Abb. 19.11). Da die erwarteten \hat{y} -Werte in beiden Nationen auf den Geraden liegen, müssen auch die adjustierten Mittelwerte auf diesen Geraden liegen. Die adjustierten Mittelwerte geben also den mittleren Unterschied zwischen beiden Nationen an, der nicht auf Unterschiede im Erleben positiver Affekte zurückgeführt werden kann. Bei der Interpretation der adjustierten Mittelwerte müssen einige Aspekte beachtet werden, die wir diskutieren, nachdem wir die Kovarianzanalyse für zentrierte unabhängige Variablen behandelt haben.

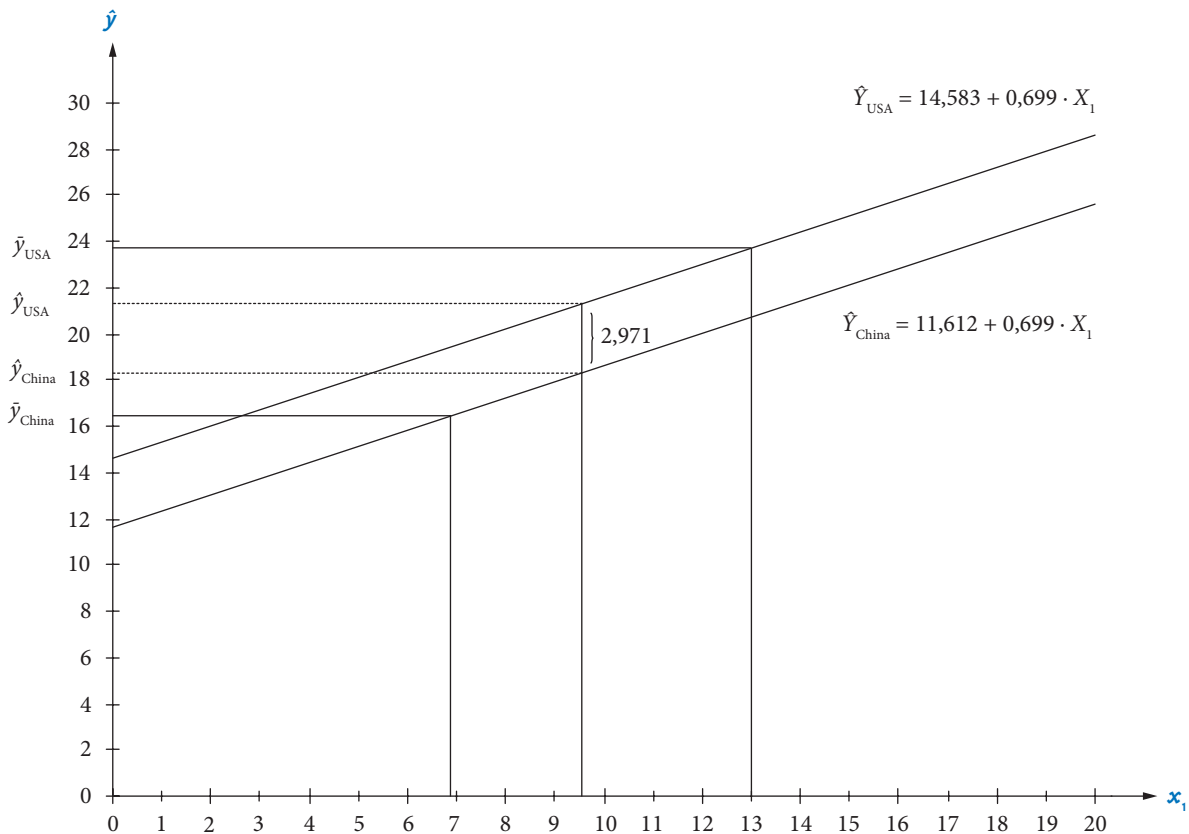


Abbildung 19.11 Lebenszufriedenheit in Abhängigkeit von dem Erleben positiver Affekte und der Nation: Adjustierte und nicht-adjustierte Mittelwerte

Zentrierte Kovariaten

Zentrierte unabhängige Variablen haben wir schon bei der moderierten Regressionsanalyse kennengelernt. Auch bei der Kovarianzanalyse hat es Vorteile, die kontinuierlichen unabhängigen Variablen (Kovariaten) vor der Analyse zu zentrieren. Ein Vorteil besteht darin, dass der Mittelwert aller Kovariaten dann gleich 0 ist. Der adjustierte Mittelwert einer Gruppe entspricht bei zentrierten Kovariaten dem Achsenabschnitt plus dem Wert der Gruppe auf der Codiervariablen. In unserem Beispiel ergeben sich für die beiden Nationen die folgenden Regressionsgleichungen:

$$\hat{y}_{\text{China}} = 18,339 + 0,699 \cdot 0 = 18,339$$

$$\hat{y}_{\text{USA}} = 18,339 + 0,699 \cdot 0 + 2,971 = 21,310$$

Man sieht, dass sich durch die Zentrierung weder die Regressionsgewichte noch die adjustierten Mittelwerte ändern. Die Achsenabschnitte nehmen zwangsläufig

ein neuen Wert an, da in ihre Bestimmung die Mittelwerte der Kovariaten einfließen (s. F 19.14). Die Zentrierung führt dazu, dass die Mittelwerte der Kovariaten sich ändern und nun gleich 0 sind. Bei zentrierten Kovariaten entspricht der Achsenabschnitt dem adjustierten Mittelwert in der Gruppe, die auf allen Codiervariablen eine 0 aufweist. In unserem Beispiel entspricht der Achsenabschnitt dem adjustierten Mittelwert in China. Das Regressionsgewicht der Dummy-Variablen gibt den Mittelwertsunterschied zwischen China und den USA wieder. Ist dieses Regressionsgewicht – wie in unserem Beispiel – signifikant von 0 verschieden (bei $\alpha = 0,05$), zeigt dies an, dass sich die adjustierten Mittelwerte bedeutsam unterscheiden. In unserem Beispiel können daher die Unterschiede der beiden Nationen im Erleben positiver Affekte den Nationenunterschied nicht vollständig erklären. Es gibt noch einen bedingten Nationenunterschied.

Interpretation der adjustierten Mittelwerte

Bei der Interpretation der adjustierten Mittelwerte sind verschiedene Aspekte zu beachten (Agresti & Finlay, 2009; Cohen et al., 2003).

Korrektheit des Regressionsmodells. Die adjustierten Mittelwerte sind nur dann angemessene Schätzwerte für die Mittelwerte in den beiden Populationen (China, USA), wenn das Regressionsmodell korrekt ist. Dies bedeutet, dass die Beziehung zwischen der Lebenszufriedenheit und dem Erleben positiver Affekte in beiden Nationen linear und das Regressionsgewicht in beiden Ländern gleich sein muss. Letztere Annahme bedeutet, dass es keine Interaktion zwischen den beiden unabhängigen Variablen (PA, Nation) gibt. Beide Annahmen können falsch sein, und ihre Gültigkeit muss überprüft werden, bevor die adjustierten Mittelwerte interpretiert werden. Wir werden in Abschnitt 19.12.3 zeigen, wie dies geschehen kann, und feststellen, dass wir in unserem Beispiel die Grundannahme der Kovarianzanalyse, dass es keine Interaktion zwischen den kategorialen unabhängigen Variablen und den Kovariaten gibt, nicht aufrechterhalten können. Wir können daher die berichteten adjustierten Mittelwerte nicht sinnvoll interpretieren. Wir haben das Beispiel bewusst gewählt, um deutlich zu machen, dass die Annahmen, die der Berechnung der adjustierten Mittelwerte zugrunde liegen, vorher überprüft werden müssen. Dies ist im Rahmen der multiplen Regressionsanalyse einfach möglich. Problematisch ist die Annahme insbesondere dann, wenn sich die Gruppen sehr stark in ihren Mittelwerten auf den Kovariaten unterscheiden und sich deren Verteilungen – im Extremfall – gar nicht überlappen. Dann lässt sich nicht überprüfen, ob in beiden Gruppen das Regressionsmodell mit gleichen Regressionsgewichten über den gesamten Wertebereich der Kovariaten hinweg gilt. Würden sich jedoch die Regressionsmodelle in ihren Steigungen in den Bereichen, die sich nicht überlappen, in der Population unterscheiden, dann käme es zu einer Fehlinterpretation der adjustierten Mittelwerte. Bei sehr großen Unterschieden in den Mittelwerten auf den Kovariaten sollten adjustierte Mittelwerte deshalb nur äußerst vorsichtig interpretiert werden.

Abhängigkeit adjustierter Mittelwertsunterschiede von den unabhängigen Variablen. Die beiden Nationen unterscheiden sich möglicherweise noch in anderen un-

abhängigen Variablen, die nicht ins Modell aufgenommen wurden. Der bedingte Nationeneffekt kann nur in Bezug auf die im Modell repräsentierten unabhängigen Variablen interpretiert werden. Es könnte sein, dass durch Hinzunahme weiterer Kovariaten, die mit der Variablen Nation korreliert sind, der bedingte Unterschied zwischen den Gruppen weiter schrumpft. Die Aufnahme aller relevanten Variablen in das Modell ist insbesondere dann wichtig, wenn kausale Interpretationen vorgenommen werden. Wir werden dies anhand eines Beispiels aus der quasi-experimentellen Evaluationsforschung im nächsten Abschnitt verdeutlichen und zeigen, dass man mit kausalen Schlussfolgerungen äußerst vorsichtig sein muss.

Messfehlerfreiheit der unabhängigen Variablen. Wir haben schon in Abschnitt 17.9.2 gesehen, dass die Regressionsanalyse voraussetzt, dass die unabhängigen Variablen fehlerfrei gemessen wurden. Dies ist in der Psychologie nur für wenige Variablen möglich. Unsere Nationenvariable dürfte dazugehören, die Variable zur Erfassung des Erlebens positiver Affekte sicherlich nicht. Wir werden die Konsequenzen des Messfehlers und Möglichkeiten, damit umzugehen, ausführlich in den Kapiteln 23–26 behandeln. Da der Messfehler jedoch gravierende Konsequenzen für die Interpretation adjustierter Mittelwerte haben kann und Unterschiede in den adjustierten Mittelwerten artifiziell erzeugen kann, werden wir die Konsequenzen des Messfehlers für die Interpretation adjustierter Mittelwerte im nächsten Abschnitt anhand eines Beispiels der quasi-experimentellen Evaluationsforschung illustrieren. Eine Konsequenz des Messfehlers ist es, dass die Regressionsgewichte verzerrt geschätzt werden, was erhebliche Konsequenzen für die Interpretation der adjustierten Mittelwerte haben kann.

19.12.2 Kovarianzanalyse in quasi-experimentellen Designs

Wir wollen einige Probleme, die mit der Interpretation adjustierter Mittelwerte in der quasi-experimentellen Forschung verbunden sind, am Beispiel eines sehr häufig realisierten Designs aufzeigen: des Interventions-Kontrollgruppen-Designs mit Vorher- und Nachhermessung. Wir werden dieses Thema sehr ausführlich behandeln, da diesem Design in der Evaluationsfor-

schung eine große Bedeutung zukommt und die Probleme, die mit der Interpretation adjustierter Mittelwerte verknüpft sind, in Lehrbüchern häufig nur knapp behandelt werden.

Wir wählen ein Beispiel aus der Bildungsforschung (s. folgenden Kasten), da man in diesem Bereich häufig nur nicht-experimentell arbeiten kann, d.h. Gruppen (z. B. Bundesländer oder Schulformen) vergleicht, die bereits gegeben sind und nicht kontrolliert manipuliert werden können. Außerdem sind die Ergebnisse von Bildungsstudien und ihre Interpretati-

onen mit weitreichenden Konsequenzen verbunden. Wir wollen die Auswirkungen aufzeigen, die die Verletzung von zwei Voraussetzungen hat: (1) Die Kovariate wurde nicht messfehlerfrei gemessen; (2) es wurden nicht alle relevanten unabhängigen Variablen in die Modellgleichung aufgenommen, d.h., es liegt ein sog. Underfitting vor. Wir arbeiten mit simulierten Daten, da wir dann die Zusammenhänge in der Population kennen und die Auswirkungen genau identifizieren können.

Beispiel

Schulform und Lesekompetenz

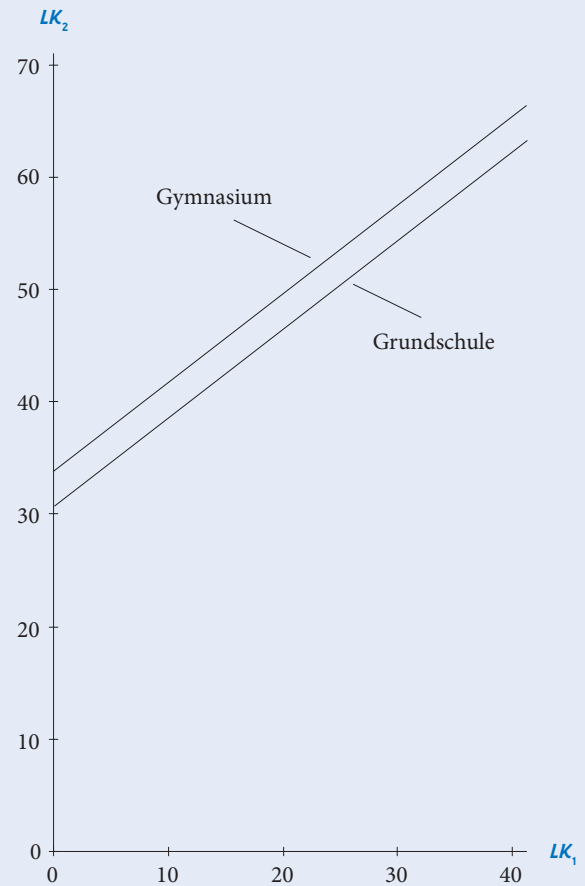
In einer europäischen Hauptstadt gibt es die Wahl, nach dem vierten Schuljahr an der Grundschule bis zum sechsten Schuljahr zu verbleiben, um dann auf eine weiterführende Schule, z. B. das Gymnasium, zu wechseln, oder bereits nach dem vierten Schuljahr den Wechsel zu vollziehen. Ein Forscher interessiert sich für die Auswirkungen eines frühzeitigen Wechsels auf die weiterführende Schule und untersucht zwei große Schulen zu Beginn des fünften und nach dem sechsten Schuljahr in Bezug auf die Lesekompetenz (LK) der teilnehmenden Schüler. Die eine Schule ist eine Grundschule (GS) mit 500 Schülern, die andere ein Gymnasium (GM) mit 500 Schülern. Zur Überprüfung von Unterschieden in der Lesekompetenz nach dem sechsten Schuljahr (abhängige Variable: LK_2) führt er eine Kovarianzanalyse durch. Dabei ist die Schulform die (kategoriale) unabhängige Variable (SF ; Grundschule: $SF = 0$, Gymnasium: $SF = 1$) und

die Lesekompetenz zu Beginn des fünften Schuljahres (LK_1) die Kovariate. Die Logik dieses Modells ist folgende: Da Kinder mit einer höheren Lesekompetenz tendenziell früher aufs Gymnasium wechseln, könnte es sein, dass Unterschiede in der Lesekompetenz am Ende des sechsten Schuljahres (LK_2) nicht wirklich auf Unterschiede in der Schulform, sondern vielmehr auf bereits vorhandene Unterschiede in der Lesekompetenz (zu Beginn des fünften Schuljahres) zurückzuführen sind. Der Forscher möchte mithilfe der Kovarianzanalyse nun überprüfen, ob die Schulform über die bereits bestehenden Unterschiede in der Lesekompetenz hinaus einen Beitrag zur Erklärung der Lesekompetenz am Ende des sechsten Schuljahres leistet. Er erhält folgendes Ergebnis einer Kovarianzanalyse:

$$\widehat{LK}_2 = 30,560^{**} + 0,797^{**} \cdot LK_1 + 2,241^{**} \cdot SF \quad (F 19,79)$$

Sowohl die Lesekompetenz zu Beginn des fünften Schuljahres als auch die Schulform haben einen bedeutsamen Einfluss. Das Regressionsgewicht der Schulform zeigt den Unterschied in den adjustierten Mittelwerten an. Im Gymnasium ist die durchschnittliche Leseleistung um den Wert 2,241 höher als in der Grundschule, wenn man den Einfluss der Lesekompetenz zu Beginn des fünften Schuljahres statistisch kontrolliert. In Abbildung 19.12, die die erwarteten LK_2 -Werte in Abhängigkeit von den LK_1 -Werten und der Schulform wiedergibt, wird dies deutlich. Das Gymnasium liegt in allen Bereichen über der Grundschule. Da insbesondere auch die guten Schüler im Gymnasium eine höhere Lesekompetenz nach dem sechsten Schuljahr aufweisen als die Schüler mit einem vergleichbaren LK_1 -Wert in der Grundschule, interpretiert unser Forscher dies als ein Beleg dafür, dass ein frühzeitiger Wechsel zum Gymnasium die Lesekompetenz steigert, und empfiehlt der Schulbehörde eine Umgestaltung des Schulsystems. Bisher erscheinen alle Ergebnisse erwartungskonform und plausibel. Sollte man dieser Empfehlung folgen?

Abbildung 19.12 Lesekompetenz nach dem sechsten Schuljahr (LK_2) in Abhängigkeit von der Lesekompetenz zu Beginn des fünften Schuljahres (LK_1) und der Schulform (fiktives Beispiel)



Um die Frage zu beantworten, schauen wir uns die Datensituation etwas genauer an. In Tabelle 19.24 sind die Mittelwerte, Standardabweichungen und Retestkorrelationen angegeben. Die Korrelationen zwischen den LK_1 -Werten und den LK_2 -Werten in den beiden Schulformen liegen bei ungefähr 0,80, was für kognitive Fähigkeiten wie die Lesekompetenz typisch ist. Die Schüler der beiden Schultypen unterscheiden sich im Mittel nach dem sechsten Schuljahr um ungefähr 10 Einheiten in der Lesekompetenz, was einer relativ großen Effektstärke entspricht. Sie unterscheiden sich – wie zu erwarten war – auch bereits zu Beginn des fünften Schuljahres um etwa 10 Punkte. Darüber hinaus fällt auf, dass der mittlere Kompetenzzuwachs in beiden Gruppen mit etwa 10 Punkten annähernd gleich groß ist. Überprüft man die Differenzwerte $LK_2 - LK_1$ zwischen den beiden Schulformen mittels eines t -Tests für unabhängige Stichproben auf Signifikanz, erhält man

einen t -Wert von $t = -0,397$; $df = 998$; $p = 0,692$. Das Ergebnis zeigt, dass der Mittelwertsunterschied auf einem vorher festgelegten Signifikanzniveau von $\alpha = 0,05$ nicht signifikant von 0 verschieden ist. Beide Schulformen unterscheiden sich somit nicht hinsichtlich des Zuwachses an Lesekompetenz. Dieses Ergebnis führt zu einer vollkommen anderen Schlussfolgerung als das Ergebnis der Kovarianzanalyse, obwohl dieselben Daten untersucht wurden. Dieses Phänomen nennt man Lords Paradox. Der Name geht auf den Psychometriker Frederic M. Lord zurück, der dieses Phänomen 1967 beschrieben hat (Lord, 1967).

Lords Paradox

Lords Paradox beschreibt das Phänomen, dass die Analyse von mittleren Veränderungen und die Kovarianzanalyse zu einander widersprechenden Ergebnissen führen, wenn Veränderungen für natürlich vorgefun-

Tabelle 19.24 Lesekompetenz in einer Grundschule und einem Gymnasium (fiktive Daten). Angegeben sind Mittelwerte und Standardabweichungen sowie Retestkorrelationen

	5. Schuljahr (Beginn)	6. Schuljahr (Ende)	Retestkorrelation
Grundschule	$\bar{x}_{GS} = 100,333$ $\hat{\sigma}_{X-GS} = 14,473$	$\bar{y}_{GS} = 110,567$ $\hat{\sigma}_{Y-GS} = 14,493$	$r_{XY-GS} = 0,804$
Gymnasium	$\bar{x}_{GM} = 110,227$ $\hat{\sigma}_{X-GM} = 15,072$	$\bar{y}_{GM} = 120,698$ $\hat{\sigma}_{Y-GM} = 15,081$	$r_{XY-GM} = 0,790$

dene, d.h. nicht randomisiert gebildete Gruppen in einem Vortest-Nachtest-Design untersucht werden. Das Paradox lässt sich in unserem Beispiel darauf zurückführen, dass die Messung der Lesekompetenz messfehlerbehaftet ist. Während der Messfehler der abhängigen Variablen (LK_2) in das Residuum der multiplen Regressionsanalyse einfließt, bleibt der Messfehler der Kovariate (LK_1) unberücksichtigt. Der Messfehler »stört« den t -Test der mittleren Veränderungen nicht, da der t -Test auf den Mittelwerten basiert und sich deshalb die individuellen Messfehler ausmitteln (s. hierzu ausführlich Abschn. 23.1.3). Dies ist bei der Kovarianzanalyse aber nicht der Fall, wie wir im Folgenden zeigen werden.

Der Messfehler und seine Auswirkungen

Um die Auswirkungen des Messfehlers zu verstehen, empfehlen wir, zunächst Kapitel 23 zu lesen. Wir werden aber versuchen, die Konsequenzen des Messfehlers so darzustellen, dass man sie auch ohne Lektüre von Kapitel 23 nachvollziehen kann. Daher werden wir zunächst die Grundidee der Messfehlertheorie zusammenfassen. Der Messfehlertheorie zufolge setzt sich ein beobachteter Wert aus einem wahren Wert (engl. true score) und einem Messfehler zusammen. Die wahren Werte sind Ausprägungen einer latenten (nicht beobachtbaren) Variablen, die True-Score-Variable genannt und mit τ bezeichnet wird. Die Messfehlervariable (kurz: Fehlervariable) wird mit ε bezeichnet. Eine beobachtete Variable lässt sich somit in eine True-Score-Variable und die Fehlervariable zerlegen. Für eine unabhängige Variable X folgt daher: $X = \tau + \varepsilon$. Ein beobachteter Lesekompetenzwert spiegelt somit z.T. die Lesekompetenz und z.T. den Messfehler wider. Aus der Zerlegung der beobachteten Variablen X folgt auch die Zerlegung ihrer Varianz $Var(X) = Var(\tau) + Var(\varepsilon)$. Wie wir in Abschnitt 23.1.4

zeigen werden, wird das Verhältnis aus der Varianz der True-Score-Variablen τ und der Varianz der beobachteten Variablen X Reliabilität genannt; es gibt an, in welchem Ausmaß beobachtete Unterschiede wahre Unterschiede widerspiegeln.

Datensimulation. Die Daten in Tabelle 19.24 wurden durch ein Simulationsverfahren erzeugt. Wir gingen hierbei von folgenden Gegebenheiten in der Population aus: Die Populationsmittelwerte für die wahren Werte betrugen in der Grundschule in der fünften Klasse 100 und in der sechsten Klasse 110, in dem Gymnasium 110 und 120. Es wurde angenommen, dass die wahren Werte im fünften Schuljahr in beiden Schulformen normalverteilt sind mit einer Standardabweichung von 13,42. Für beide Schulformen wurden aus dieser Verteilung 500 wahre Werte gezogen. Der wahre Wert eines Schülers im Nachtest (LK_2) ergab sich als sein Vortestwert (LK_1) plus einer Konstanten von 10. Jeder Schüler hat sich also in Bezug auf seine wahren Werte in exakt gleichem Ausmaß verändert, unabhängig davon, welcher Schule er angehört. Zu jedem Vortestwert und jedem Nachtestwert wurde ein Fehlerwert addiert, und zwar derart, dass die Reliabilitäten 0,80 betragen.

Erklärung der Auswirkungen des Messfehlers. Bevor wir die Auswirkungen des Messfehlers formal erläutern, wollen wir zunächst zeigen, dass sich die Berücksichtigung des Messfehlers auch wirklich auf die Ergebnisse der Analysen in unserem Beispiel auswirkt. Berechnet man die Kovarianzanalyse nicht auf der Grundlage der messfehlerbehafteten beobachteten Vortestwerte, sondern der wahren Vortestwerte der einzelnen Schüler (τ_{LK_1}), erhält man folgende Gleichung:

$$\widehat{LK}_2 = 11,067^{**} + 0,993^{**} \cdot \tau_{LK_1} - 0,166 \cdot SF \text{ (F 19.80)}$$

Diese Gleichung zeigt uns zwei wichtige Ergebnisse:

- (1) Der Einfluss der Schulform ist in dieser Gleichung nicht mehr signifikant. Die Hypothese, dass die adjustierten Mittelwerte in beiden Schulformen in der Population gleich sind, wird nicht verworfen. Der hoch signifikante Effekt der Kovariate hingegen war theoretisch zu erwarten: Die Unterschiede im Nachtest sind v. a. darauf zurückzuführen, dass sich die Schüler im Vortest unterscheiden und die beiden Schulformen unterschiedliche Vortestmittelwerte aufweisen.
- (2) Das Regressionsgewicht der Kovariate (also der Vortestwerte) ist größer, wenn man die wahren LK_1 -Werte betrachtet, als wenn man die beobachteten (messfehlerbehafteten) LK_1 -Werte betrachtet. Hierin liegt des Rätsels Lösung. Durch den Messfehler wird das Regressionsgewicht der Kovariate im Modell mit den beobachteten LK_1 -Werten unterschätzt. Dadurch fällt die Adjustierung der LK_2 -Werte durch die LK_1 -Werte schwächer aus, als es aufgrund der wahren Werte notwendig gewesen wäre. Damit wird auch die Differenz in den erwarteten Nachtestwerten zwischen beiden Schulformen unterschätzt: Da sich beide Schulformen in ihren mittleren Ausgangswerten unterscheiden, ist der erwartete Unterschied in den Nachtestwerten bei einem Regressionsgewicht von 0,797 (ohne Berücksichtigung des Messfehlers) geringer als bei einem Regressionsgewicht von 0,993 (mit Berücksichtigung des Messfehlers). Dies wird dann durch das signifikante Regressionsgewicht der Variablen, die die Schulform repräsentiert, ausgeglichen.

Man kann das Phänomen wie folgt veranschaulichen. In der Population wurde der wahre Nachtestwert einer Person bestimmt, indem zum Vortestwert der Wert 10 hinzuaddiert wurde. Daher gilt auch für die Mittelwerte in den beiden Schulformen auf Populationsebene:

$$\text{Mittelwert}(\text{Nachtest}_{\text{GM}}) = 10 + \text{Mittelwert}(\text{Vortest}_{\text{GM}})$$

$$\text{Mittelwert}(\text{Nachtest}_{\text{GS}}) = 10 + \text{Mittelwert}(\text{Vortest}_{\text{GS}})$$

Hieraus ergibt sich für die Mittelwärtsdifferenz im Nachtest:

$$\begin{aligned} \text{Mittelwert}(\text{Nachtest}_{\text{GM}}) - \text{Mittelwert}(\text{Nachtest}_{\text{GS}}) \\ = \text{Mittelwert}(\text{Vortest}_{\text{GM}}) - \text{Mittelwert}(\text{Vortest}_{\text{GS}}) \end{aligned}$$

Diese Beziehung bleibt erhalten, wenn wir uns die Ergebnisse der Regressionsanalyse mit der fehlerbereinigten Kovariate in der Stichprobe anschauen. Das Regressionsgewicht ist (aufgrund des Stichprobenfehlers) nur unwesentlich von 1 verschieden, die erwartete Mittelwärtsdifferenz im Nachtest gleich der Mittelwärtsdifferenz im Vortest, wie es den Daten in Tabelle 19.24 – abgesehen vom Stichprobenfehler – auch entspricht.

Verzerre Schätzung des Regressionsgewichts durch Messfehler. Der Messfehler beeinflusst nicht die Mittelwerte der Variablen, da sich die Messfehler ausmitteln. In den Stichprobendaten sind daher die Mittelwärtsunterschiede im Nachtest – abgesehen vom Stichprobenfehler – gleich den Mittelwärtsunterschieden im Vortest. Allerdings beeinflusst der Messfehler die Berechnung des Regressionsgewichts. Betrachten wir hierzu die beiden Schulformen getrennt. Innerhalb einer Schulform ist das Regressionsgewicht gleich dem Regressionsgewicht der einfachen linearen Regression. Dies wird nach Gleichung F 17.4 wie folgt bestimmt:

$$b_1 = \frac{s_{XY}}{s_X^2}$$

In der Population dann entsprechend:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Wie wir in Abschnitt 23.2.1 zeigen werden, wird in der Messfehlertheorie typischerweise angenommen, dass die Messfehler verschiedener Variablen voneinander unabhängig sind, da die Messfehler zufällige, unsystematische Einflüsse widerspiegeln. Daher folgt, dass die Kovarianz zweier Variablen X und Y gleich der Kovarianz ihrer True-Score-Variablen τ_X und τ_Y ist: $\text{Cov}(X, Y) = \text{Cov}(\tau_X, \tau_Y)$. Für die Varianz von X ergibt sich: $\text{Var}(X) = \text{Var}(\tau_X) + \text{Var}(\varepsilon_X)$ und somit für das Regressionsgewicht:

$$\beta_1 = \frac{\text{Cov}(\tau_X, \tau_Y)}{\text{Var}(\tau_X) + \text{Var}(\varepsilon_X)}$$

Diese Formel zeigt zwei wichtige Sachverhalte: (1) Das Regressionsgewicht der messfehlerbehafteten Variablen hängt nur vom Messfehler der unabhängigen Variablen X , nicht aber vom Messfehler der abhängigen Variablen Y ab. (2) Es wird umso kleiner, je größer die Fehlervarianz und damit je geringer die Reliabilität ist. Dies lässt sich für das Regressionsgewicht eines Modells mit einer

unabhängigen Variablen einfach herleiten. Das Regressionsgewicht der multiplen Regressionsanalyse wird in komplexer Weise vom Messfehler der unabhängigen Variablen beeinflusst. Es kann sowohl zu Unter- als auch Überschätzungen kommen. Für unsere Argumentation reicht es jedoch aus zu sehen, dass das Regressionsgewicht durch Messfehlereinflüsse auf die unabhängigen Variablen verzerrt wird. In unserem Populationsmodell ist ein einfaches lineares Modell das korrekte Modell zur Vorhersage der Lesekompetenz nach dem sechsten Schuljahr. Würden wir nun aufgrund der messfehlerbehafteten Variablen in der Gesamtstichprobe eine einfache Regressionsanalyse durchführen, bekämen wir folgende Regressionsgleichung:

$$\widehat{LK}_2 = 29,272^{**} + 0,820^{**} \cdot LK_1 \quad (F\ 19,81)$$

Aufgrund des Messfehlers wird das Regressionsgewicht erwartungsgemäß unterschätzt. Es beträgt 0,820 und nicht 1. Für die beiden Schulformen ergeben sich hieraus folgende erwarteten Mittelwerte von LK_2 :

$$\begin{aligned} \text{erwarteter Mittelwert(Nachtest}_{GM}) \\ = 29,272 + 0,820 \cdot \text{Mittelwert(Vortest}_{GM}) \end{aligned}$$

$$\begin{aligned} \text{erwarteter Mittelwert(Nachtest}_{GS}) \\ = 29,272 + 0,820 \cdot \text{Mittelwert(Vortest}_{GS}) \end{aligned}$$

Hieraus ergibt sich die erwartete Mittelwertsdifferenz:

$$\begin{aligned} \text{Mittelwert(Nachtest}_{GM}) - \text{Mittelwert(Nachtest}_{GS}) \\ = 0,820 \cdot [\text{Mittelwert(Vortest}_{GM}) \\ - \text{Mittelwert(Vortest}_{GS})] \end{aligned}$$

Man erwartet aufgrund des verzerrt geschätzten Regressionsgewichts, dass die Nachtestdifferenz nur das 0,820-fache der Vortestdifferenz beträgt. Wie wir Tabelle 19.24 entnehmen, sind beide Differenzen jedoch annähernd gleich. Um das Regressionsmodell optimal an die Daten anzupassen, nutzt die Regressionsanalyse die Möglichkeit, die sich bietet, wenn die Schulform als kategoriale Variable zusätzlich in die Regressionsgleichung aufgenommen wird. Das Regressionsgewicht der Schulform gleicht die durch das verzerrte Regressionsgewicht des Vortests erwartete zu geringe Mittelwertsdifferenz des Nachtests aus, indem das Regressionsgewicht, das ja ebenfalls die Differenz zwischen beiden Schulformen (Ausprägungen der Dummy-Variablen) gewichtet, einen positiven Wert erhält.

Wichtige Erkenntnisse. Aus den bisherigen Ausführungen ergeben sich sechs wichtige Erkenntnisse:

- (1) Der Messfehler der unabhängigen Variablen kann zu Fehlinterpretationen des Effekts der kategorialen Variablen führen. In unserem Beispiel wurde der Effekt der Schulform überschätzt; er kann aber auch unterschätzt werden (s. hierzu ausführlich Campbell & Kenny, 1999, Kap. 5).
- (2) Die Berücksichtigung des Messfehlers behebt das Problem, das durch die Messfehlerverzerrung entsteht.
- (3) Im Falle messfehlerbehafteter unabhängiger Variablen sollte zu Modellen mit latenten Variablen (s. Kap. 23–26) gegriffen werden, um die Verzerrung zu beheben.
- (4) Es sollte angestrebt werden, unabhängige Variablen – soweit es geht – ohne Messfehler zu messen.
- (5) Kausalinterpretationen (hier: Effekt der Schulform) sollten in der quasi- und nicht-experimentellen Forschung mit großer Vorsicht vorgenommen werden.
- (6) Im Gegensatz zur Kovarianzanalyse führt der t -Test der Differenzwerte zum korrekten Schluss, auch ohne dass die Variablen in Bezug auf ihren Messfehler korrigiert wurden. Dies erklärt sich daraus, dass der t -Test auf einem Mittelwertsvergleich basiert und die Messfehler bei der Mittelung der Werte ausgemittelt werden.

Kovarianzanalyse und messwiederholte Varianzanalyse

Die Daten in Tabelle 19.24 könnte man auch mit einer zweifaktoriellen Varianzanalyse mit Messwiederholung auf einem Faktor auswerten (s. Abschn. 14.2.2). Man hätte dann einen zweistufigen Gruppenfaktor (Schulform) und einen zweistufigen Messwiederholungsfaktor (Messzeitpunkte: Beginn der fünften Klasse, Ende der sechsten Klasse). In unserem Beispiel würde man einen Haupteffekt für die Schulform erwarten, da die Gymnasialschüler auf beiden Kompetenzmessungen höhere Werte haben. Man würde außerdem einen Haupteffekt der Messzeitpunkte erwarten, da die Lesekompetenz generell zunimmt. Man würde aber keinen Interaktionseffekt erwarten, da sich die Kinder in beiden Schulformen gleich entwickeln. Genau dies ist auch der Fall. Der F -Wert der Interaktion beträgt $F = 0,158$ ($df_1 = 1$, $df_2 = 998$, $p = 0,692$) und ist nicht signifikant. Der F -Wert ist exakt gleich dem quadrierten t -Wert des t -Tests für die Differenzwerte, beide Tests führen

zum selben Ergebnis und selben p -Wert. Zur Analyse differenzieller Veränderungen in Abhängigkeit von der Schulform können daher sowohl der t -Test für Differenzwerte als auch die zweifaktorielle Varianzanalyse mit Messwiederholung auf einem Faktor herangezogen werden.

Auswirkungen ausgelassener relevanter unabhängiger Variablen

Kommen wir zum zweiten Problem, dem Problem ausgelassener relevanter unabhängiger Variablen. Die Kovarianzanalyse führt in der nicht-experimentellen Forschung nur dann zu einem validen Schluss in Be-

zug auf den Effekt einer unabhängigen (kategorialen) Variablen, wenn keine relevanten Variablen ausgelassen werden. Der Effekt der unabhängigen Variablen wird auch dann verzerrt geschätzt, wenn es weitere Variablen gibt, die mit der unabhängigen Variablen zusammenhängen und über die im Modell bereits berücksichtigten Variablen hinaus die abhängige Variable beeinflussen würden – hätte man sie ins Modell aufgenommen. Werden sie ausgelassen, wird der Effekt falsch geschätzt, und diese Verzerrung lässt sich nicht einfach wie beim Messfehler durch die Messfehlerbereinigung ausgleichen. Auch dies wollen wir an unserem Beispiel illustrieren.

Beispiel

Schulvergleich in der Lesekompetenz: Effekte ausgelassener unabhängiger Variablen

Wir haben in einer Erweiterung unserer Simulationsstudie angenommen, dass innerhalb der beiden Schulformen der wahre Nachtestwert nicht nur von dem Vortestwert abhängt, sondern darüber hinaus von einer zweiten Variablen, die wir soziale Schicht (SOS) nennen und die wir aus Einfachheitsgründen dichotom mit den Werten 0 (niedrige soziale Schicht) und 1 (hohe soziale Schicht) erfasst haben. Innerhalb jeder Schulform liegt folgende Gleichung der Datensimulation zugrunde:

$$\text{Wahrer } LK_2\text{-Wert} = 10 + \text{wahrer } LK_1\text{-Wert} + 5 \cdot \text{SOS}$$

Bei gleichen wahren Vortestwerten haben Kinder aus einer höheren sozialen Schicht im Vergleich zu Kindern aus einer niedrigeren sozialen Schicht einen um 5 erhöhten Nachtestwert. Man könnte sich diesen Unterschied dadurch erklären, dass in Familien aus einer höheren sozialen Schicht mehr Bücher vorhanden sind, mehr Leseanreize gesetzt werden, mehr mit den Kindern gelesen wird usw. Wichtig ist, dass innerhalb jeder Schulform die Gleichung dieselbe ist. Kinder innerhalb der gleichen sozialen Schicht entwickeln sich also in beiden Schulformen gleich. Wir haben nun die Daten so erzeugt, dass die soziale Schicht mit der Schulform korreliert ist, und zwar derart, dass der Anteil hoher sozialer Schicht im Gymnasium 80% und bei den in der Grundschule verbliebenen Schülern 20% beträgt. Das Populationsmodell lautet also für die Gesamtpopulation:

$$\begin{aligned} \text{Wahrer } LK_2\text{-Wert} &= 10 + \text{wahrer } LK_1\text{-Wert} \\ &+ 5 \cdot \text{SOS} + 0 \cdot \text{SF} \end{aligned}$$

Die Schulform ist mit der sozialen Schicht korreliert, hat aber über sie hinaus keinen Effekt. Der Unterschied zwischen den beiden Schulformen ist also nicht auf den Unterricht zurückzuführen, sondern auf die ungleiche Verteilung der Variablen SOS über die beiden Schulformen hinweg. Sowohl zu dem wahren Vortestwert als auch zu dem wahren Nachtestwert wurde ein Fehlerwert addiert. Die beobachteten Variablen wurden einer Kovarianzanalyse unterzogen. Die soziale Schicht wurde zunächst nicht berücksichtigt. Man erhält folgendes Ergebnis:

$$\widehat{LK}_2 = 31,786^{**} + 0,795^{**} \cdot LK_1 + 5,355^{**} \cdot \text{SF} \quad (\text{F } 19.82)$$

Man sieht, dass sowohl das Regressionsgewicht der Kovariate (Vortest; LK_1) als auch das Regressionsgewicht der Schulform signifikant von 0 verschieden sind. Der Schuleffekt ist gegenüber Gleichung F 19.79 noch größer geworden. Man wäre folglich noch eher verführt, das Ergebnis dieser Analyse dahin gehend zu deuten, dass der frühzeitige Wechsel zum Gymnasium mit einem deutlich höheren Kompetenzgewinn einhergeht als das Verbleiben an der Grundschule. Was passiert, wenn man anstatt der fehlerbehafteten beobachteten Werte für LK_1 die wahren Werte im Vortest in die Regressionsgleichung aufnimmt? Man erhält in diesem Fall folgende Gleichung:

$$\widehat{LK}_2 = 12,609^{**} + 0,987^{**} \cdot \tau_{LK_1} + 2,982^{**} \cdot \text{SF} \quad (\text{F } 19.83)$$

Das Regressionsgewicht der Kovariate wird nun – abgesehen vom Stichprobenfehler – korrekt geschätzt. Dadurch wird auch der Effekt der Schulform verringert, da sie nicht mehr die Unterschätzung der Mittelwerte aufgrund des verzerrten Regressionsgewichts ausgleichen muss. Aber der Effekt der Schulform verschwindet nicht, sondern bleibt bedeutsam. Dies ist darauf zurückzuführen, dass die Schulform mit der sozialen Schicht zusammenhängt, die nicht in die Regressionsgleichung aufgenommen wurde. Der Effekt der Schulform ist also ein Scheineffekt.

Was passiert nun, wenn wir die soziale Schicht mit in die Regressionsgleichung aufnehmen und zunächst die messfehlerbehaftete Kovariate berücksichtigen? Wir erhalten die Gleichung:

$$\widehat{LK}_2 = 30,767^{**} + 0,797^{**} \cdot LK_1 + 2,766^{**} \cdot SF + 4,157^{**} \cdot SOS \quad (\text{F 19.84})$$

Gegenüber Gleichung F 19.82 hat sich das Gewicht der Schulform verringert, ist aber weiterhin signifikant. Es ist ein bedeutsamer Effekt der sozialen Schicht hinzugekommen. Erst wenn wir die Kovariate um ihren Messfehler bereinigen und die True-Score-Variable in das Modell aufnehmen, sind die Abweichungen der

geschätzten Parameter von den Populationsparametern nur auf den Stichprobenfehler zurückzuführen, und der Effekt der Schulform verschwindet:

$$\widehat{LK}_2 = 11,111^{**} + 0,993^{**} \cdot \tau_{LK_1} - 0,078 \cdot SF + 4,860^{**} \cdot SOS \quad (\text{F 19.85})$$

In diesem Beispiel wäre auch die zweifaktorielle Varianzanalyse mit Messwiederholung auf einem Faktor nicht in der Lage, den wahren Effekt aufzudecken, wenn die soziale Schicht nicht mit in das Modell aufgenommen wird. Dies rührt daher, dass sich die beiden Schulformen in ihren mittleren Zuwächsen bedeutsam unterscheiden (s. Tab. 19.25). Dieser Unterschied zeigt jedoch keinen Schuleffekt an, sondern geht auf die Konfundierung der Schulform mit der sozialen Schicht zurück. Erst wenn man die Varianzanalyse um den Faktor SOS erweitert, verschwindet auch dort der Effekt, dass die Schulform mit dem Messwiederholungsfaktor interagiert; dies zeigt, dass sich beide Schulformen nicht in ihren bedingten Effekten unterscheiden. Für die adäquate Interpretation des Schuleffekts ist es daher notwendig, dass alle relevanten Variablen in die Regressionsgleichung aufgenommen werden.

Tabelle 19.25 Lesekompetenz in einer Grundschule und einem Gymnasium (fiktive Daten)

a Mittelwerte und Standardabweichungen

	5. Schuljahr (Beginn)	6. Schuljahr (Ende)
Grundschule	$\bar{x}_{GS} = 100,333$ $\hat{\sigma}_{X-GS} = 14,473$	$\bar{y}_{GS} = 111,547$ $\hat{\sigma}_{Y-GS} = 14,595$
Gymnasium	$\bar{x}_{GM} = 110,227$ $\hat{\sigma}_{X-GM} = 15,072$	$\bar{y}_{GM} = 124,768$ $\hat{\sigma}_{Y-GM} = 15,096$

b Korrelationen zwischen den Variablen Lesekompetenz (LK), Schulform (SF) und soziale Schicht (SOS)

	LK ₁	LK ₂	SF
LK₂	0,814**		
SF	0,318**	0,407**	
SOS	0,183**	0,320**	0,618**

** $p \leq 0,01$

Schlussfolgerungen

Die beiden Simulationsstudien haben gezeigt, dass Fehlinterpretationen in Bezug auf den Effekt der kategorialen Variablen in der nicht-experimentellen Forschung sowohl durch die Messfehlerabhängigkeit der Kovariaten als auch durch das Auslassen relevanter Variablen bedingt sein können. Da die Kovarianzanalyse ein Spezialfall der multiplen Regressionsanalyse ist, trifft dies auch allgemein auf die Regressionsanalyse zu.

Allgemein gilt, dass Regressionskoeffizienten verzerrt geschätzt werden, wenn die unabhängigen Variablen mit einem Messfehler gemessen wurden. Der Messfehler der abhängigen Variablen vergrößert die Residualwerte, die Messfehlervarianz vergrößert die Residualvarianz. Dies hat zur Folge, dass der Determinationskoeffizient unterschätzt wird. Der Messfehler der abhängigen Variablen wirkt sich aber nicht verzerrend auf die Schätzung der Regressionskoeffizienten

aus. Vielmehr vergrößert er die Vorhersageungenauigkeit. Die Regressionskoeffizienten werden im Allgemeinen auch verzerrt geschätzt, wenn relevante unabhängige Variablen nicht in die Regressionsgleichung aufgenommen wurden.

Konsequenzen für die Anwendung der Regressionsanalyse. Was bedeutet dies für die Anwendung der multiplen Regressionsanalyse? Man muss darauf achten, dass der Messfehlereinfluss bei den unabhängigen Variablen relativ gering ist, und man muss möglichst alle relevanten Variablen in das Modell aufnehmen. Im Falle der Messfehlerbehaftetheit bietet es sich an, auf moderne Verfahren mit latenten Variablen auszuweichen (s. Kap. 26). Bei sehr reliablen Messinstrumenten ist die Verschätzung der Regressionskoeffizienten relativ gering. Sehr viel schwieriger ist es, einen Kausaleffekt abzusichern. Dies geht streng genommen nur, wenn man Personen zu den Werten der unabhängigen Variablen per Zufall zuweist (Randomisierung, s. Abschn. 4.3.3). Ist dies nicht durchführbar, müssen möglichst alle relevanten Störvariablen identifiziert und in die Modellgleichung aufgenommen werden. Um die Regressionsgewichte im Sinne kausaler Effekte interpretieren zu können, müssen also besondere Maßnahmen der Kontrolle von konfundierten Variablen ergriffen werden. Mit spezifischen Verfahren zur Absicherung von Kausalinterpretationen in der psychologischen Forschung beschäftigen sich z. B. die Arbeiten von Rubin (2006) sowie Steyer (1992, 2003). Baumert et al. (2009) zeigen anhand einer empirischen Untersuchung zum Frühübergang in ein Gymnasium, wie solche Methoden gewinnbringend eingesetzt werden können, um Fehlinterpretationen zu vermeiden.

Auch wenn man die konfundierten Variablen nicht alle kennt und berücksichtigen kann, ist es in verschiedenen Forschungskontexten sinnvoll, die Regressionsanalyse einzusetzen. Man darf die Effekte dann nur nicht leichtfertig im Sinne kausaler Effekte interpretieren. In unserem Beispiel kann es durchaus sinnvoll sein, die Variable Schulform in die Regressionsanalyse aufzunehmen, auch wenn man die soziale Schicht nicht kennt. Dies ist z. B. dann sinnvoll, wenn man die Lesekompetenz nach dem sechsten Schuljahr präzisieren möchte. Für diesen Zweck leistet die Schulform einen bedeutsamen Beitrag, wenn man die soziale Schicht nicht kennt oder nicht berücksichtigt.

Man muss sich dann aber bewusst sein, dass es sich nicht um den wahren, erklärenden kausalen Effekt handelt.

19.12.3 Interaktionen zwischen kategorialen und kontinuierlichen Variablen

Die Überprüfung von Interaktionen folgt dem bereits in den Abschnitten zur moderierten Regression und zu kategorialen unabhängigen Variablen beschriebenen Vorgehen (s. Abschn. 19.9 und 19.11). Die metrischen unabhängigen Variablen werden mit den Codiervariablen multipliziert, und die so entstandenen Produktvariablen werden in die Regressionsgleichung aufgenommen. Signifikanzstatistisch wird überprüft, ob diese Interaktionsterme zusätzlich zu den metrischen unabhängigen Variablen und den Codiervariablen einen bedeutsamen Varianzanteil des Kriteriums erklären. In unserem Beispiel aus der kulturvergleichenden Lebenszufriedenheitsforschung haben wir die Interaktionsvariable zwischen der Nation und der Häufigkeit positiver Affekte berechnet, indem wir die beiden Variablen miteinander multipliziert und diese Interaktionsvariable zusätzlich in die Regressionsgleichung aufgenommen haben. Die Analyse ergibt folgende Gleichung, wobei in Klammern wieder die Standardfehler angegeben sind:

$$\begin{aligned}\hat{Y} = & 13,329^{**} + 0,452^{**} \cdot X_1 - 2,111^{*} \cdot X_2 \\ & (0,495) \quad (0,063) \quad (1,008) \\ & + 0,506^{**} \cdot X_1 \cdot X_2 \\ & (0,090)\end{aligned}$$

Die Variable X_1 kennzeichnet das Erleben positiver Affekte, die Variable X_2 die Nation. Die Analyse zeigt, dass der Einfluss der Interaktionsvariablen bedeutsam ist. In diesem Beispiel wurde die metrische Variable nicht zentriert, woraus sich der vergleichsweise hohe Standardfehler des Regressionsgewichts der Variablen X_2 erklärt, die zu $r = 0,95$ mit der Produktvariablen korreliert ist. Wir haben in diesem Anwendungsfall die Variable nicht zentriert, um die Auswirkung der Interaktion der Variablen in Bezug auf die ursprüngliche Metrik der Variablen in Abbildung 19.13 grafisch darstellen zu können. Darüber hinaus hat der Nullpunkt der Variablen X_1 eine klare Bedeutung. Die Interaktionsvariable klärt zusätzlich zu den anderen Varia-

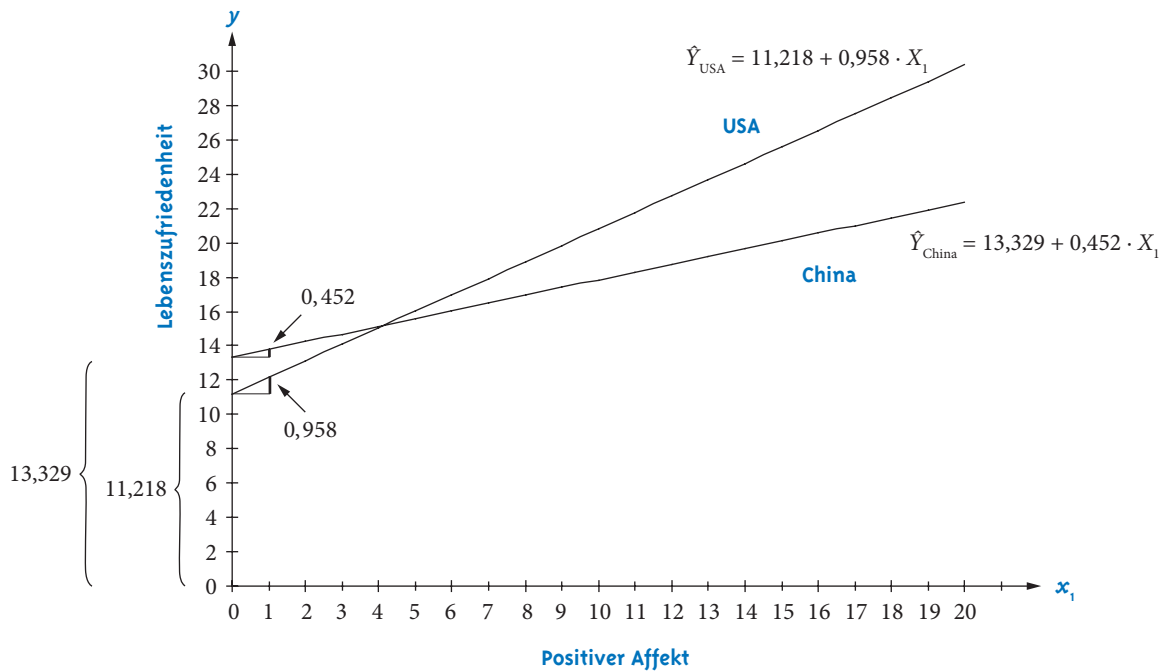


Abbildung 19.13 Lebenszufriedenheit in Abhängigkeit von dem Erleben positiver Affekte in zwei verschiedenen Nationen: Interaktion zwischen den unabhängigen Variablen

blen 1,9% Varianz auf. Die weitere Hinzunahme des Quadrates der Variablen PA (X_1) sowie des Produktes aus der quadrierten PA-Variablen und der Nationenvariablen (X_2) führt zu keiner signifikanten Erhöhung des Determinationskoeffizienten R^2 (R^2 -Änderung = 0,003; $F = 2,143$, $df_1 = 2$, $df_2 = 992$; $p = 0,118$). Der Zusammenhang zwischen der Häufigkeit positiver Affekte und der Lebenszufriedenheit innerhalb der beiden Nationen ist also nicht kurvilinear.

Auch wenn der Zuwachs an Varianzaufklärung durch die Produktvariable auf den ersten Blick gering erscheint, sieht man sowohl in Abbildung 19.13 als auch in Tabelle 19.26, dass die Interaktion bedeutsame Auswirkungen auf die erwarteten Werte der Lebenszufriedenheit hat und die Missachtung der Interaktion bedeutsame Prognosefehler nach sich ziehen würde.

In Abbildung 19.13 sind die Regressionsgeraden dargestellt, die man für beide Nationen erhält, wenn man deren Werte in die Regressionsgleichung einsetzt. Die Grafik zeigt, dass sich beide Geraden schneiden. Im unteren Bereich der Skala zur Erfassung des Erlebens positiver Affekte ist die Lebenszufriedenheit in China größer, während sie im hohen Wertebereich in den USA größer ist. Die Regressionsgerade ist in den USA

steiler als in China, d. h., die Lebenszufriedenheit hängt in den USA stärker vom Erleben positiver Affekte ab als in China. Suh et al. (1998) geben hierzu eine ausführliche theoretische Begründung, die v. a. auf der Bedeutung von positiven Gefühlen in individualistischen (USA) vs. kollektivistischen (China) Nationen beruht.

Konsequenzen ausgelassener Interaktionsvariablen

Um zu veranschaulichen, zu welchen Fehlprognosen es kommen kann, wenn wichtige Variablen nicht in die Regressionsanalyse aufgenommen wurden, sind in Tabelle 19.26 vier Regressionsmodelle und ihre Konsequenzen für die Prognose von Lebenszufriedenheit in beiden Nationen dargestellt. Im ersten Regressionsmodell wird nur der Einfluss des Erlebens positiver Affekte untersucht, im zweiten Regressionsmodell wird ausschließlich die Nation als unabhängige Variable aufgenommen. Im dritten Regressionsmodell, das dem Modell der Kovarianzanalyse entspricht, werden die Effekte des Erlebens positiver Affekte und der Nation simultan geschätzt. Das vierte Modell enthält zusätzlich die Interaktion zwischen Erleben positiver Affekte und Nation. In diesen vier verschiedenen Regressionsgleichungen wurde jeweils ein Wert von 20 für die

Häufigkeit des Erlebens positiver Affekte eingesetzt, den ein virtueller Herr Li in China und ein virtueller Herr Miller in den USA aufweisen könnten. Anhand der Regressionsgleichung wird dann der erwartete Lebenszufriedenheitswert berechnet. Man sieht, dass die vier Regressionsmodelle zu unterschiedlichen Ergebnissen kommen. Insbesondere der Vergleich des Modells, das die Interaktion enthält, mit dem Modell, das der Kovarianzanalyse entspricht, ist in Bezug auf die Bewertung der Bedeutung der Interaktion relevant. Hier sieht man, dass bei Missachtung der Interaktion die Lebenszufriedenheitswerte der beiden Personen deutlich überschätzt werden: Die Lebenszufriedenheit von Herrn Li wird in dem ausschließlich additiven Modell überschätzt, die Lebenszufriedenheit von Herrn Miller unterschätzt. Die Hinzunahme der Interaktion führt zu einer präziseren Prognose der Lebenszufriedenheit beider Personen.

Aptitude-Treatment-Interaction-Analyse

Ein Regressionsmodell, das eine metrische unabhängige Variable, eine kategoriale unabhängige Variable sowie deren Interaktion enthält, wird in der Fachliteratur auch unter dem Begriff der Aptitude-Treatment-Interaction-Analyse diskutiert (Snow, 1989). Dieser Begriff stammt aus der Pädagogischen Psychologie. Anhand eines entsprechenden Designs wird untersucht, ob sich verschiedene Interventionen (Treatments) in Abhängigkeit von der Ausgangsleistung oder Begabung (Aptitude) unterschiedlich auswirken. Verschiedene Untersuchungen haben gezeigt, dass der Effekt eines Treatments für Personen unterschiedlicher Ausgangsleistung unterschiedlich ausfällt. Solche Fragestellungen können im Rahmen der multiplen Regressionsanalyse angemessen untersucht werden.

19.13 Regressionsdiagnostik

Wir haben in den verschiedenen Abschnitten dieses Kapitels dargelegt, auf welchen Annahmen die multiple Regressionsanalyse aufbaut und was alles beachtet werden muss, wenn man eine multiple Regressionsanalyse durchführt. In diesem Abschnitt werden wir verschiedene Möglichkeiten vorstellen, um zu überprüfen, ob diese Annahmen erfüllt sind und ob es auffällige Strukturen in den Daten gibt, die die Interpretation gefährden. Wir behandeln folgende Themen:

- (1) korrekte Spezifikation des Modells
- (2) Messfehlerfreiheit der unabhängigen Variablen
- (3) Ausreißer und einflussreiche Datenpunkte
- (4) Multikollinearität
- (5) Homoskedastizität
- (6) Unabhängigkeit der Residuen
- (7) Normalverteilung der Residuen

19.13.1 Korrekte Spezifikation des Modells

Eine wichtige Voraussetzung für die korrekte Interpretation der Ergebnisse ist, dass das Modell korrekt spezifiziert wird. Damit ist gemeint, dass keine relevanten Variablen ausgelassen wurden und dass Variablen, die in der Population keinen bedeutsamen Beitrag zur Vorhersage oder Erklärung der abhängigen Variablen leisten, nicht ins Modell aufgenommen wurden. Das Auslassen relevanter Variablen nennt man auch *Underfitting*, die Aufnahme von irrelevanten Variablen auch *Overfitting*. Wir haben Konsequenzen des Under- und Overfitting an verschiedenen Stellen dieses Kapitels besprochen und wollen die Konsequenzen nun zusammenfassend diskutieren.

Tabelle 19.26 Erwartete Lebenszufriedenheitswerte für die fiktiven Personen Herrn Li und Herrn Miller, die beide über einen Wert von 20 auf der Variablen zur Erfassung positiver Affekte (X_1) verfügen, in vier Regressionsmodellen

	Nur PA (X_1)	Nur Nation (X_2) 0: China, 1: USA	Additiv	Interaktiv
Regressionsgleichung	$\hat{Y} = 11,122 + 0,887 \cdot X_1$	$\hat{Y} = 16,464 + 7,207 \cdot X_2$	$\hat{Y} = 11,612 + 0,699 \cdot X_1 + 2,971 \cdot X_2$	$\hat{Y} = 13,329 + 0,452 \cdot X_1 - 2,111 \cdot X_2 + 0,506 \cdot X_1 \cdot X_2$
Herr Li	28,862	16,464	25,592	22,369
Herr Miller	28,862	23,671	28,563	30,378

Underfitting

Kurvilinearität. Die Beziehung zwischen der abhängigen Variablen und den unabhängigen Variablen ist nicht linearer, sondern kurvilinear Art. Hätten wir in dem Beispiel in Abschnitt 19.10, in dem wir die Zufriedenheit mit einem Kurs auf die wahrgenommenen Anforderungen zurückführten, die quadratische Komponente nicht in das Regressionsmodell aufgenommen, wären wir zu einem vollkommen falschen Schluss über den Zusammenhang beider Variablen gekommen. Während beide Variablen im rein linearen Modell nicht bedeutsam zusammenhängen, ist der Zusammenhang im quadratischen Regressionsmodell sehr hoch und signifikant von 0 verschieden.

Interaktionen. In unserem kulturvergleichenden Beispiel in Abschnitt 19.12.3 würde man den Zusammenhang zwischen der Lebenszufriedenheit und dem Erleben positiver Affekte falsch einschätzen und die Lebenszufriedenheitswerte verzerrt prognostizieren, wenn man die Produktvariable, die die Interaktion zwischen beiden Variablen repräsentiert, auslassen würde.

Konfundierte Variablen. In unserem hypothetischen Schulbeispiel in Abschnitt 19.12.2 haben wir gesehen, wie man das bedeutsame Regressionsgewicht der Schulform im Sinne eines Schuleffektes fehldeuten kann, wenn nicht alle Variablen, die mit der Schulform korreliert (konfundiert) sind, in die Regressionsgleichung aufgenommen wurden. Die kausale Interpretation von Regressionsgewichten in nicht-experimentellen Studien, die mit der multiplen Regressionsanalyse ausgewertet werden, kann nur mit äußerster Vorsicht vorgenommen werden und setzt neben der zeitlichen Vorgeordnetheit der verursachenden Variablen voraus, dass alle konfundierten Variablen in die Regressionsgleichung aufgenommen und kontrolliert wurden. Ein Regressionsgewicht kann daher nur in Bezug auf die im Modell enthaltenen anderen unabhängigen Variablen interpretiert werden.

Aufdecken von Underfitting. Wie kann man Underfitting aufdecken? Die Berücksichtigung der relevanten konfundierten Variablen setzt typischerweise voraus, dass man eine theoretische Vorstellung von dem Wirkprozess hat und die potenziellen Störvariablen bei der Planung der Studie erfasst hat. In unserem Schulbeispiel

ist es also wichtig, all diejenigen Variablen aufgrund theoretischer Überlegungen schon *vor* der Datenerhebung theoretisch zu identifizieren und zu erheben, die mit der Schulform potenziell korreliert sind. Diese sollten bereits vor dem Schulwechsel erhoben werden, um ihren Einfluss auf die Schulwahl und die abhängige Variable angemessen untersuchen zu können. Häufig wird man aber nicht alle potenziellen Störvariablen kennen, sodass man mit der kausalen Interpretation entsprechend vorsichtig sein muss.

Ob relevante Produktvariablen ausgelassen wurden, kann überprüft werden, indem diese in die Regressionsgleichung aufgenommen werden und untersucht wird, ob sie einen bedeutsamen Einfluss haben. In ähnlicher Weise kann auch vorgegangen werden, um die Nicht-Linearität eines Effekts zu überprüfen. Hat man theoretische Vorstellungen von der Form des nicht-linearen Zusammenhangs, kann ein entsprechendes Regressionsmodell spezifiziert und überprüft werden. Hierzu können Polynome höherer Ordnung gebildet werden und getestet werden, ob diese einen bedeutsamen Einfluss haben. Auch können andere nicht-lineare Regressionsmodelle spezifiziert werden. Statistikprogramme bieten hierzu verschiedene Modelle an wie z.B. die logarithmische Regression oder die exponentielle Funktion.

Hat man keine theoretischen Vorstellungen von der Form des Zusammenhangs, kann man sich auch anhand der Daten eine Funktion anpassen lassen. Eine Möglichkeit besteht darin, sich eine LOWESS-(oder LOESS-)Anpassungslinie angeben zu lassen. Diese Linie folgt keiner Regressionsgleichung und zählt daher zu den nonparametrischen Regressionsverfahren. Sie erlaubt es, die Daten in Bezug auf die Art des Zusammenhangs zwischen beiden Variablen exploratorisch zu analysieren.

LOWESS-Anpassungsverfahren. Durch ein LOWESS-Anpassungsverfahren kann eine Linie in das Punktediagramm eingepasst werden, die den Zusammenhang zwischen beiden Variablen widerspiegelt, ohne dass eine konkrete Gleichung angegeben werden muss. Bei diesem Verfahren handelt es sich um ein Glättungsverfahren (*LO*cally *WE*ighted Scatterplot Smoother; Cleveland, 1979). Die Methode arbeitet in groben Zügen wie folgt: Für einen Wert x_m einer unabhängigen Variablen X betrachtet man eine vorher festzulegende Anzahl von x -Werten um diesen Wert

herum. Meistens bemisst man die Anzahl der betrachteten x -Werte anhand eines Häufigkeitskriteriums (z. B. 10 % aller Werte). Für diese x -Werte bestimmt man eine lineare Regression und ordnet dem Zielwert x_m denjenigen erwarteten \hat{y} -Wert zu, der auf der Regressionsgeraden liegt. Dies macht man nun für alle x -Werte und enthält somit die erwarteten \hat{y} -Werte. Je nachdem, wie viele Werte man um einen Zielwert x_m zulässt, ist die angepasste Linie stärker oder weniger stark geglättet. Durch Ausprobieren kann man diejenige Linie auswählen, die die Daten angemessen repräsentiert. Die LOWESS-Anpassungslinie kann man z. B. mit dem Computerprogramm SPSS unter der Option »Scatterplot« erstellen. In Abbildung 19.14 ist eine solche Anpassungslinie dargestellt, wobei 70 % der Werte um einen x -Wert berücksichtigt wurden. Sie zeigt, dass der Zusammenhang kurvilinear von Natur ist. Die Anpassung einer quadratischen Regression ist daher sinnvoll. Die parametrische Regression hat den Vorteil, dass die Regressionsgleichung zur Vorhersage individueller \hat{y} -Werte genutzt werden kann. Als exploratives Verfahren kann aber die LOWESS-Anpassung gute Dienste leisten. Solche Glättungsverfahren lassen sich auch für die Regression mit mehreren unabhängigen Variablen und für robuste Verfahren nutzen (s. hierzu Wilcox, 2003).

Verringerte Teststärke. Hat man nicht alle relevanten unabhängigen Variablen berücksichtigt, so hat dies auch zur Folge, dass die erklärte Varianz geringer ist.

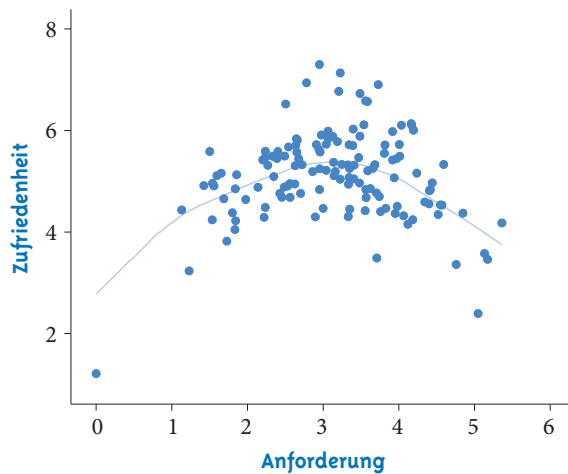


Abbildung 19.14 LOWESS-Anpassungslinie für den Zusammenhang zwischen Zufriedenheit mit einem Kurs und wahrgenommener Anforderung

Da man hierdurch an Teststärke verliert, benötigt man eine größere Stichprobe, um einen Effekt inferenzstatistisch abzusichern. Den Verlust an Teststärke erkennt man daran, dass der Determinationskoeffizient kleiner ist, wenn nicht alle relevanten unabhängigen Variablen berücksichtigt wurden. Hierdurch vergrößert sich der Nenner in der Prüfgröße F 19.32.

Overfitting

Konsequenzen des Overfitting haben wir schon in Abschnitt 19.7.7 kennengelernt. Die Hinzunahme irrelevanter unabhängiger Variablen kann zu einer verzerrten Schätzung der Regressionsgewichte der anderen Variablen führen, wodurch Prognose- und Kreuzvalidierungsfehler begünstigt werden. Irrelevante unabhängige Variablen sollten daher aus der Regression genommen werden.

19.13.2 Messfehlerfreiheit der unabhängigen Variablen

Wir haben den Einfluss des Messfehlers ausführlich am Beispiel des Zusammenhangs von Schulform und Lesekompetenz im Rahmen der Kovarianzanalyse in Abschnitt 19.12.2 behandelt. In der einfachen Regressionsanalyse führt der Messfehler der unabhängigen Variablen zu einer Unterschätzung des wahren Regressionsgewichts, in einer multiplen Regressionsanalyse können die Verzerrungen in unterschiedliche Richtungen gehen. Wie wir am Beispiel des Zusammenhangs von Schulform und Lesekompetenz gesehen haben, kann der Messfehler bei der Erfassung der Lesekompetenz sogar zu einer Verzerrung des Regressionsgewichts einer Variablen (Schulform) führen, die messfehlerfrei gemessen wurde. Der Grad der Messfehlerabhängigkeit der Messungen der unabhängigen Variablen sollte daher bestimmt werden. Hierzu kann man auf die Reliabilität zurückgreifen. Wir werden in Abschnitt 23.2 zeigen, wie man diese schätzen kann. Sind die unabhängigen Variablen in hohem Maße messfehlerbehaftet, sollte auf Regressionsmodelle mit latenten Variablen zurückgegriffen werden, die wir in Kapitel 26 behandeln werden.

19.13.3 Ausreißer und einflussreiche Datenpunkte

Wir haben schon bei der Produkt-Moment-Korrelation in Abschnitt 16.3.1 gesehen, dass die Korrelation durch einzelne Ausreißer sehr verzerrt werden kann. Da die Regressionsanalyse eng mit der Korrelationsanalyse verwandt ist, können Ausreißer auch die Ergebnisse einer Regressionsanalyse verzerren. Ausreißer sind Werte, die sich stark von den restlichen Werten unterscheiden. Ausreißerwerte können sowohl auf der abhängigen als auch auf den unabhängigen Variablen auftreten. Wir werden daher zeigen, wie man Ausreißerwerte auf beiden Variablenarten identifizieren kann. Ausreißerwerte können, müssen aber nicht zwangsläufig die Parameterschätzung verzerren. Wenn es z. B. für einen festen x -Wert einen positiven und einen negativen Ausreißerwert auf der Y -Variablen gibt, muss dies keinen Effekt auf die Schätzung der Regressionsgeraden haben. Es wurde daher zusätzlich noch das Konzept der einflussreichen Datenpunkte (engl. influential data points) entwickelt. Ein einflussreicher Datenpunkt ist dadurch gekennzeichnet, dass sich die Schätzungen der Regressionsparameter und der vorhergesagten Werte stark verändern, wenn dieser Datenpunkt (z. B. die Wertekombination einer Person) den Daten entnommen wird.

Zur Aufdeckung von Ausreißerwerten und einflussreichen Datenpunkten gibt es statistische Kennwerte, die wir im Folgenden beschreiben werden. Über die Bestimmung dieser Kennwerte hinaus sollte man die Daten vor einer Regressionsanalyse immer genau kontrollieren.

Datenkontrolle

Bei der Datenkontrolle ist es wichtig zu überprüfen, ob alle Daten innerhalb der Wertebereiche der Variablen liegen oder ob es Daten gibt, die allein schon durch ihren Wert darauf hinweisen, dass ein Eingabefehler vorliegen muss. Wir erinnern uns noch an eine Kollegin aus Amerika, die ihre Studierenden nach deren Gewicht gefragt hat und – ohne dass die Studierenden bei der Befragung darüber informiert waren – beim Verlassen des Hörsaals das Gewicht aller Studierenden durch eine Waage bestimmt hat. Die Korrelation zwischen dem selbst eingeschätzten und dem objektiv gemessenen Gewicht lag nahe bei 0. Dieses Ergebnis war spektakulär – spricht es doch für eine starke Selbsttäuschung. Allerdings zeigte sich bei der Datenkontrolle,

dass die Kollegin vergessen hatte, dem Statistikprogramm mitzuteilen, dass alle fehlenden objektiv gemessenen Gewichtswerte mit 99 codiert waren, sodass das Programm den Wert 99 als wahren Gewichtswert interpretierte. Nach Behebung dieses Fehlers korrelierte das objektive Gewicht mit dem selbst eingeschätzten Gewicht zu ungefähr $r = 0,80$. Eine Person mit dem Wert 99 wäre ein einflussreicher Datenpunkt gewesen. Dieses Beispiel zeigt, wie wichtig die Datenkontrolle ist. Hierauf ist schon bei der Dateneingabe zu achten. Bei Fragebogenstudien erlebt man z. B. immer wieder, dass Personen, die die Untersuchung boykottieren wollen, absurde Antwortmuster produzieren. Werden diese nicht erkannt und die entsprechenden Personen nicht aus dem Datensatz entfernt, kann es zu schwerwiegenden Ergebnisverfälschungen kommen.

Identifikation von Ausreißerwerten auf den unabhängigen Variablen

Um Ausreißerwerte auf einer unabhängigen Variablen aufzudecken, kann man auf die Mahalanobis-Distanz und die Hebelwerte zurückgreifen. Wir illustrieren diese Kennwerte nur für den Fall einer einfachen Regressionsanalyse, da sie im multiplen Fall matrixalgebraisch bestimmt werden müssen. Die Übertragung auf den multiplen Fall ist konzeptuell jedoch einfach. Die Kennwerte geben jeweils an, wie stark die Werte der Personen auf den unabhängigen Variablen von ihrem Mittelwert abweichen.

Mahalanobis-Distanz. Ein extremer Wert liegt vor, wenn die Abweichung des individuellen Wertes vom Mittelwert groß ist. Da die Richtung der Abweichung keine Rolle spielt, quadriert man den Abweichungswert. Da die Größe der Abweichung vom Maßstab abhängt, teilt man den quadrierten Abweichungswert durch die Varianz. Um diese Abweichung auf die ursprüngliche Metrik zu beziehen, berechnet man die Quadratwurzel aus diesem Quotienten und erhält für jede Person m ihren Mahalanobis-Distanzwert d_m , benannt nach dem indischen Statistiker Prasanta Chandra Mahalanobis (1893–1972):

$$d_m = \sqrt{\frac{(x_m - \bar{x})^2}{\hat{\sigma}_x^2}} \quad (\text{F 19.86})$$

Statistikprogramme wie SPSS berechnen typischerweise den quadrierten Mahalanobis-Distanzwert d_m^2

und berichten diesen als Mahalanobis-Distanzwert. Je größer dieser Wert ist, umso stärker ist die Abweichung. Die Werte in Tabelle 19.27 zeigen für das Beispiel des Zusammenhangs zwischen der studentischen Zufriedenheit mit einem Kurs und den wahrgenommenen Anforderungen in diesem Kurs, dass die quadrierten Mahalanobis-Distanzwerte zwischen 0,445 und 53,668 liegen. Die Mahalanobis-Distanzwerte lassen sich direkt in Hebelwerte überführen, die wir im Folgenden vorstellen und für die wir kritische Werte angeben werden.

Hebelwert. Der Begriff Hebelwert (engl. leverage = Hebelwirkung) rührt daher, dass Werte der unabhängigen Variablen, die sehr weit von ihrem Mittelwert entfernt sind, sich stärker auf die Bestimmung der Regressionsgewichte auswirken. Dies kann man sich anhand von Abbildung 17.4 im Kapitel zur einfachen Regressionsanalyse veranschaulichen (s. Abschn. 17.9.6). Wir haben anhand dieser Abbildung gesehen, dass das Konfidenzintervall für die Regressionsgerade von der Ausprägung der unabhängigen Variablen abhängt: Je weiter ein Wert der unabhängigen Variablen von seinem Mittelwert abweicht, umso größer ist das Konfidenzintervall der Regressionsgeraden. Man stelle sich nun die Regressionsgerade als Stab im Raum vor. Würde man die Regressionsgerade am Mittelpunkt der x -Achse verankern und mit konstanter Kraft an ihr ziehen, dann wäre der Ausschlag – die Hebelwirkung – umso größer, je weiter entfernt vom Ankerpunkt man ziehen würde. Die Hebelwerte zeigen diese Hebelwirkung an.

Wir wollen den statistischen Hintergrund ihrer Bestimmung nicht beleuchten (s. hierzu z. B. Stevens, 2009) und die Hebelwerte nur für den Fall einer einzigen unabhängigen Variablen formal angeben. In diesem Fall berechnet sich der Hebelwert einer Person zu:

$$h_m = \frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1) \cdot \hat{\sigma}_X^2} = \frac{1}{n} + \frac{d_m^2}{n-1} \quad (\text{F 19.87})$$

Zieht man von den Hebelwerten den Wert $1/n$ ab, erhält man die zentrierten Hebelwerte. Diese entsprechen der quadrierten Mahalanobis-Distanz geteilt durch $(n-1)$. Die zentrierten Hebelwerte können Werte zwischen 0 und $1 - 1/n$ annehmen, in unserem Beispiel also Werte zwischen 0 und 0,993. Zur Bewertung dieser Hebelwerte werden in der Literatur verschiedene Schwellenwerte diskutiert, ab denen zentrierte Hebelwerte als auffällig gelten. So schlagen z. B. Belsey et al. (2004) als unteren Schwellenwert $2 \cdot k/n$ bei großen und $3 \cdot k/n$ bei kleinen Stichproben vor. In unserem Beispiel wären Schwellenwerte von $2 \cdot k/n = 2 \cdot 2/136 = 0,029$ und $3 \cdot k/n = 3 \cdot 2/136 = 0,044$ in Betracht zu ziehen. Über der ersten Schwelle liegen in unserem Beispiel 9,6 % der Werte, über der zweiten Schwelle 5,88 % der Werte. Die Schwellenwerte wurden ausgehend von der Überlegung bestimmt, bei normalverteilten Variablen ungefähr die 5 % der extremsten Werte aufzudecken. Cohen et al. (2003) weisen darauf hin, dass man anhand dieser Schwellen meist zu viele auffällige Werte erhält, und empfehlen, sich die Verteilung der Hebelwerte anzuschauen und nur diejenigen Werte genauer

Tabelle 19.27 Zufriedenheit und wahrgenommene Anforderung: Identifikation von Ausreißern: Residuenstatistik, Mahalanobis-Distanz und zentrierter Hebelwert ($n = 136$)

	Minimum	Maximum	Mittelwert	Standardabweichung
Nicht standardisierte Residuen	−1,840	1,807	0,000	0,664
Standardisierte Residuen	−2,750	2,702	0,000	0,993
Studentisierte Residuen	−2,765	2,718	0,000	1,003
Gelöschte Residuen	−1,860	1,828	−0,001	0,678
Studentisierte ausgeschlossene Residuen	−2,838	2,786	0,001	1,011
Mahalanobis-Distanz	0,445	53,668	1,985	5,115
Zentrierter Hebelwert	0,003	0,398	0,015	0,038

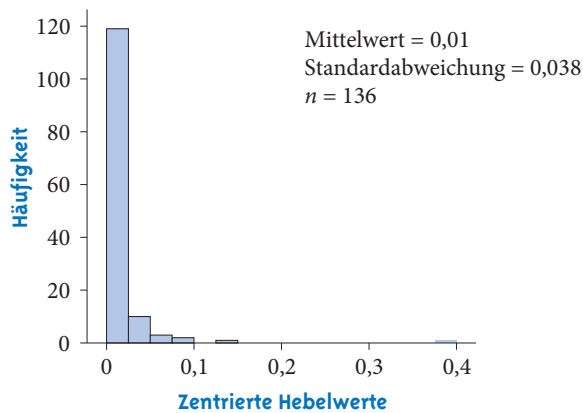


Abbildung 19.15 Histogramm der zentrierten Hebelwerte für das Beispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung«

zu inspizieren, die sich stark von den anderen unterscheiden. Abbildung 19.15 zeigt das Histogramm der Hebelwerte. Es ist ein Hebelwert auffällig, der mit 0,398 den höchsten Wert aufweist (s. auch Tab. 19.27). Dem Datensatz entnehmen wir, dass dies die Person ist, die den geringsten Wert auf der Variablen Anforderung aufweist. In unserer Studierendenstichprobe ist also die Person, die sich sehr unterfordert fühlt, auffällig, und ihr kommt eine große Hebelwirkung zu.

Identifikation von Ausreißerwerten auf der abhängigen Variablen

Bei der Identifikation von Ausreißerwerten auf der abhängigen Variablen greift man auf die Residuen zurück. In Tabelle 19.27 ist die Residuenstatistik, die von dem Statistikprogramm SPSS ausgegeben wird, zusammengestellt. Andere Computerprogramme stellen diese Information in ähnlicher Weise zur Verfügung. Zunächst werden die Residualwerte (nicht-standardisierte Residuen) der 136 Personen zusammengefasst. Der Residualwert ist – wie wir schon gesehen haben – die Differenz zwischen dem beobachteten y_m -Wert einer Person m und ihrem geschätzten \hat{y}_m -Wert. Der kleinste in unserem Datenbeispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung« gefundene Residualwert beträgt $-1,840$, der größte $1,807$, der Mittelwert der Residualwerte muss 0 sein, die Standardabweichung gibt die Streuung der Residualwerte an. Je stärker der Residualwert von 0 abweicht, umso stärker weicht der y -Wert einer Person von der Regressionsgeraden ab. Da die Größe der Residuen vom Maßstab abhängt, mit der

die Variable gemessen wurde, werden die Residuen normiert. Hierzu gibt es verschiedene Möglichkeiten, wobei deren Benennung uneinheitlich ist (z. B. Kockelkorn, 2000). Wir folgen der Namensgebung der SPSS-Ausgabe, nennen jeweils aber auch alternative Bezeichnungen, die angemessener und in der Statistik gebräuchlich sind.

Standardisiertes Residuum. Das standardisierte Residuum erhält man, indem das Residuum durch seine geschätzte Standardabweichung geteilt wird. Die geschätzte Standardabweichung ist der geschätzte Standardschätzfehler, den man nach Gleichung F 19.29 erhält. In unserem Beispiel ist die geschätzte Standardabweichung $\hat{\sigma} = 0,669$. Sie darf nicht mit der Standardabweichung der Residuen in Tabelle 19.27 verwechselt werden. Die Standardabweichung der Residuen ist die Standardabweichung der Residualwerte, die man in einer konkreten Stichprobe findet. Die geschätzte Standardabweichung ist die geschätzte Standardabweichung der Residuen in der Population.

Studentisiertes Residuum. Bei der Bestimmung des studentisierten Residuums macht man sich eine Eigenschaft des Regressionsmodells zunutze. Es lässt sich zeigen, dass die Präzision, mit der die Residuen geschätzt werden, umso größer ist, je weiter ein x -Wert von seinem Mittelwert entfernt ist. Beim studentisierten Residuum wird das Residuum nicht durch den allgemeinen geschätzten Standardschätzfehler geteilt, sondern durch die geschätzte Standardabweichung des Residuums an der Stelle x_m . Diese Standardabweichung erhält man, indem man den geschätzten Standardschätzfehler mit dem Wert $\sqrt{1 - h_m}$ multipliziert, wobei h_m den Hebelwert bezeichnet. Dieses Residuum wird auch *internally studentized residual* genannt. Es ist eine genauere Schätzung eines normierten Residuums als das standardisierte Residuum und sollte diesem vorgezogen werden.

Gelöschtes Residuum. Dieses Residuum erhält man, indem man für die Schätzung der Regressionsparameter die Person aus den Daten nimmt und ihren vorhergesagten Wert auf der Grundlage dieser Schätzgleichung bestimmt. Dies hat den Vorteil, dass die Regressionsgleichung selbst nicht schon durch die Person verzerrt wird. Hierdurch wird verhindert, dass eine Person, die einen Ausreißerwert aufweist, die Schätzgleichung schon so verändert, dass sie als Ausreißer schwerer zu identifizieren ist.

Studentisiertes ausgeschlossenes Residuum. Die Bestimmung dieser Residuen folgt demselben Prinzip wie die Bestimmung des gelöschten Residuums. Die betrachtete Person wird bei der Schätzung der Regressionsparameter nicht berücksichtigt. Auf der Grundlage dieser Regressionsparameter wird ihr vorhergesagter Wert und ihr Residualwert bestimmt. Dieser wird dann durch die geschätzte Residualstandardabweichung an der Stelle ihrer x -Werte geteilt, wobei ihr Wert in die Bestimmung dieser Standardabweichung nicht eingeflossen ist. Dieses Residuum wird in der Literatur *externally studentized residual* genannt. Manche Autoren nennen es auch einfach studentisiertes Residuum (Kockelkorn, 2000). Dieses Residuum ist das Residuum, das man der Bewertung der Residuen zugrunde legen sollte. Ist das Regressionsmodell in der Population gültig, folgt dieses Residuum einer t -Verteilung mit $df = n - k - 1$ Freiheitsgraden. Anhand der kritischen Werte der t -Verteilung kann man beurteilen, wie extrem der Residualwert ist und festlegen, welche Werte man sich genauer anschaut. Häufig wird empfohlen, ein studentisiertes Residuum genauer zu betrachten, das einen absoluten Wert aufweist, der größer als 3 ist. Dies ist in unserem Beispiel nicht der Fall.

Umgang mit Ausreißerwerten

Die in diesem Abschnitt behandelten Statistiken helfen, Ausreißerwerte zu identifizieren. Wie sollte man mit ihnen umgehen? Zunächst einmal kann ein Ausreißerwert ein ganz normaler Wert der Verteilung sein. Wenn ein Merkmal z. B. in einer Population normalverteilt ist, kann es vorkommen, dass ein extremer Wert in die Stichprobe gelangt, auch wenn dies selten ist. Ausreißer sollte man nur dann aus den Daten nehmen, wenn man einen guten Grund dafür hat. Wenn man z. B. bei einer Exploration der Ausreißerwerte merkt, dass die Wertekombination der Person auf Eingabefehler, Missverständnisse, Boykott etc. schließen lässt, dann ist dies ein guter Grund, den Ausreißerwert aus den Daten zu nehmen. Hat man solche Hinweise nicht, ist die Entscheidung schwieriger. Wirkt sich der Ausreißer wenig auf die Schätzung der Regressionsparameter aus, so kann man ihn in den Daten belassen. Wirkt er sich stark aus, so gibt es mehrere Möglichkeiten, die wir behandeln werden, nachdem wir Möglichkeiten vorgestellt haben, wie man solche einflussreichen Datenpunkte identifiziert. Die Analyse der Ausreißerstatistiken für unser Beispiel zeigt nur einen auffälligen Wert, nämlich die

Person, die sich von dem Kurs sehr unterfordert fühlte. Allerdings ist ihr Wert ein zulässiger – wenn auch seltener – Wert, und man würde den Wert daher nicht aus dem Datensatz entnehmen.

Einflussreiche Datenpunkte

Einflussreiche Datenpunkte sind dadurch gekennzeichnet, dass sich der Wert der geschätzten Regressionsgewichte stark ändert, wenn dieser Datenpunkt (in unserem Beispiel eine Person) aus den Daten herausgenommen wird. Die Veränderung kann für jeden Regressionskoeffizienten getrennt betrachtet werden. Darüber hinaus kann auch insgesamt betrachtet werden, wie stark sich die vorhergesagten \hat{y} -Werte ändern.

Änderung der Regressionskoeffizienten (DfBETA und DfBETAS)

Zur Bewertung der Änderung der Regressionskoeffizienten, die man erhält, wenn ein Datenpunkt (eine Beobachtungseinheit wie z. B. eine Person) aus dem Datensatz entfernt wird, kann man die DfBETA-Werte bestimmen. Das »Df« steht für die Differenz, »BETA« für einen Regressionskoeffizienten. Ein DfBETA-Wert ist einfach die Differenz aus dem geschätzten Regressionskoeffizienten mit und ohne die Beobachtungseinheit (z. B. Person) in der Stichprobe. Man erhält somit einen solchen Wert für jede Beobachtungseinheit und jeden Regressionskoeffizienten. In Tabelle 19.28 sind diese Werte für den Achsenabschnitt, das Regressionsgewicht der zentrierten Anforderungsvariablen und der quadrierten zentrierten Anforderungsvariablen (Anforderung²) in Bezug auf ihren Minimal-, Maximal- und Mittelwert sowie die Standardabweichung zusammengestellt. Man sieht, dass die Werte relativ klein sind.

Um die Veränderungen in Bezug auf verschiedene Regressionskoeffizienten vergleichbar zu machen, werden DfBETAS berechnet. Das »S« am Ende des Namens zeigt an, dass es sich um standardisierte Werte handelt. Man erhält sie, indem man den DfBETA-Wert durch den Standardfehler des Regressionskoeffizienten teilt, den man auf der Grundlage der Stichprobe ohne die Beobachtungseinheit (z. B. Person) berechnet. Nach Cohen et al. (2003) zeigen DfBETAS-Werte, die in kleinen bzw. mittelgroßen Stichproben vom Betrag her größer als 1 sind, auffällige Werte an; für große Stichproben empfehlen die Autoren, DfBETAS-Werte, die vom Betrag her größer als $2/\sqrt{n}$ sind, kritisch zu betrachten. In unserem Beispiel weist nach dem ersten Kriterium

Tabelle 19.28 Einflussreiche Datenpunkte: Geschätzte Werte für das Beispiel der Regression der Zufriedenheit mit einem Kurs auf die wahrgenommene Anforderung (zentrierte Anforderung)

	Minimum	Maximum	Mittelwert	Standardabweichung
DfBETA Achsenabschnitt	−0,017	0,020	0,000	0,006
DfBETA Anforderung	−0,029	0,023	0,000	0,005
DfBETA Anforderung ²	−0,022	0,022	−0,000	0,004
DfBETAS Achsenabschnitt	−0,252	0,295	0,000	0,088
DfBETAS Anforderung	−0,477	0,386	0,000	0,088
DfBETAS Anforderung ²	−0,498	0,481	−0,001	0,086
DfFIT	−0,151	0,147	−0,001	0,025
DfFITS	−0,677	0,594	−0,001	0,150
Cook-Distanz	0,000	0,148	0,007	0,017

keine Person einen extremen DfBETAS-Wert auf, nach dem zweiten Kriterium gibt es für den Achsenabschnitt neun und die beiden Regressionsgewichte jeweils sechs auffällige Werte. Die eine Person mit den beiden höchsten positiven DfBETAS-Werten für die Regressionsgewichte der linearen und der quadratischen Komponente hat einen zentrierten Anforderungswert von 2,22 und einen Zufriedenheitswert von 4,18. Die Person mit den höchsten negativen DfBETAS-Werten hat einen zentrierten Anforderungswert von 1,91 und einen Zufriedenheitswert von 2,40. Beide Personen sind in Abbildung 19.10 b (in Abschn. 19.10) am rechten Ende der Anforderungsverteilung zu finden. Sie unterscheiden sich bei ungefähr gleichem Anforderungswert in ihrem Zufriedenheitswert. Beide Personen haben jedoch plausible Wertekombinationen. Darüber hinaus ist die unstandardisierte numerische Veränderung durch die Entnahme einer der beiden Personen gering, sodass man die Personen im Datensatz belassen würde.

Änderung der vorhergesagten \hat{y} -Werte (DfFIT und DfFITS). Während sich die DfBETA-Werte auf einzelne Regressionskoeffizienten beziehen, kann man sich auch die Frage stellen, wie sich die erwarteten \hat{y} -Werte ändern, wenn man eine Person der Stichprobe entnimmt. Hierzu kann man auf die DfFIT- und die DfFITS-Werte zurückgreifen. Ein DfFIT-Wert ist die Differenz aus dem vorhergesagten Wert, den man für eine Person erhält, wenn man ihn anhand der Regressionskoeffizienten bestimmt, die man an der Gesamtstichprobe gewonnen hat, und

dem Wert, den man anhand der Regressionskoeffizienten vorhersagt, die man anhand der Stichprobe geschätzt hat, aus der die Person entnommen wurde. Teilt man diese Differenz durch den geschätzten Standardfehler der vorhergesagten Werte auf der Grundlage der ohne die Person gewonnenen Regressionskoeffizienten, erhält man das standardisierte DfFITS-Maß. Nach Cohen et al. (2003) sind DfFITS-Werte auffällig, die in kleinen bzw. mittleren Stichproben einen absoluten Betragswert aufweisen, der größer als 1 ist; für große Stichproben sehen sie einen Wert von $2 \cdot \sqrt{(k+1)/n}$ als kritische Schwelle an. In unserem Beispiel beträgt letzterer Schwellenwert 0,297. Acht Personen unserer Stichprobe weisen einen Wert auf, dessen Absolutbetrag größer als 0,297 ist. Allerdings weisen nur die beiden schon über DfBETAS identifizierten Personen mit 0,594 und −0,677 DfFITS-Werte auf, die sehr deutlich über der Schwelle liegen.

Es lässt sich zeigen, dass der DfFITS-Wert einer Person aus dem Produkt ihres studentisierten ausgeschlossenen Residuums und dem Faktor $\sqrt{h_m/(1-h_m)}$ gebildet wird. Es fließen also sowohl extreme Werte auf der abhängigen als auch solche auf der unabhängigen Variablen ein. Eine Person ist somit dann besonders einflussreich, wenn sie sowohl auf der abhängigen als auch auf der unabhängigen Variablen extreme Werte aufweist.

Cooks Distanz. Cook (1979) hat einen Koeffizienten zur Identifikation von einflussreichen Werten entwickelt, der eng mit dem DfFITS-Wert verwandt ist. Es lässt sich zeigen, dass Cooks Distanzwert eine Funkti-

on des quadrierten DfFITS-Wert ist. Dieser wird (1) mit der geschätzten Residualvarianz im Modell ohne die Person multipliziert und (2) durch die geschätzte Residualvarianz im Modell mit der Person dividiert, die mit $k + 1$ gewichtet wird. Beide Maße lassen sich somit direkt ineinander umrechnen. Wenn sich beide geschätzten Residualvarianzen nicht unterscheiden, wäre Cooks Distanz einfach der quadrierte, durch $k + 1$ dividierte DfFITS-Wert. Cooks Distanz kann keine negativen Werte annehmen. Cooks Distanz hat eine komplexe Verteilung (s. Kockelkorn, 2000). Cook empfiehlt daher aus Einfachheitsgründen, auf die Quantile der F -Verteilung mit $df_1 = k + 1$ und $df_2 = n - k - 1$ zurückzugreifen. Als kritische Schwelle empfehlen Cohen et al. (2003) das $\alpha = 0,50$ -Quantil. In unserem Fall beträgt dieses $F(3,133) = 0,793$. Nach diesem Kriterium ist keiner der Werte einflussreich (s. Tab. 19.28). Obwohl die DfFITS-Werte und Cooks Distanzwerte ineinander überführbar sind, kommen wir anhand der Schwellenwerte zu unterschiedlichen Ergebnissen, wenn wir die Schwellenwerte für große Stichproben betrachten. Legen wir die Empfehlungen für mittelgroße Stichproben zugrunde, zu der wir die unsrige zählen, kommen wir zu einer einheitlichen Bewertung derart, dass es keine einflussreichen Datenpunkte gibt.

Umgang mit einflussreichen Datenpunkten

Einflussreiche Datenpunkte können die geschätzten Regressionskoeffizienten verzerren. Dies kann gravierende Folgen haben, wenn man eine Regressionsgleichung etwa zur Prognose zukünftigen Verhaltens einsetzen will. Wenn z. B. eine Unternehmensberatungsfirma eine Regressionsgleichung zur Bestimmung des zukünftigen Berufserfolgs anhand einer Stichprobe geschätzt hat, möchte sie, dass das zukünftige Verhalten präzise vorhergesagt wird. Wie geht man mit einflussreichen Datenpunkten um? Wie bei den Ausreißerwerten ist zunächst zu klären, ob diese Werte auf Fehler zurückzuführen sind, die anzeigen, dass den Werten nicht vertraut werden kann. Ist dies der Fall, kann der Wert entnommen werden. Ist dies nicht der Fall, muss man sich überlegen, wie man weiter vorgeht. Hat man Hinweise darauf, dass für die Person andere Prozesse gelten als für den Rest der Stichprobe und dies extrem selten ist, würde man die Person herausnehmen, dies mitteilen und die Einzigartigkeit der Person beschreiben. Ist z. B. in die Stichprobe zur Bestimmung des Berufserfolgs Bill Gates gerutscht, und

bemisst man den Berufserfolg am Einkommen, könnte die Regressionsgleichung sehr verzerrt werden. Nimmt man Bill Gates heraus mit der Begründung, dass für Milliardäre andere Prozesse gelten, die man in einer eigenen Milliardärsstudie untersuchen sollte, wird jeder nachvollziehen können, dass man es vorzieht, eine valide Prognosegleichung für die Population der Nicht-Milliardäre zu schätzen und diese für die Personalauswahl zu verwenden. In diesem Beispiel hätte man eine sog. Mischverteilung (Nussbeck et al., 2010; Rost & Eid, 2009): Die Population setzt sich aus verschiedenen Subpopulationen zusammen, für die ein unterschiedliches Regressionsmodell gilt. Um solche Subpopulationen mit einem Mischverteilungsmodell aufdecken zu können, müssten die Subpopulationen entsprechend groß in der Stichprobe vertreten sein. Ein einzelner Bill Gates würde nicht ausreichen.

Einflussreiche Werte können auch auf eine Fehlspezifikation des Modells hinweisen. Entfernen wir in unserem Beispiel die quadratische Komponente aus dem Regressionsmodell, erhalten wir DfFITS- und Distanzwerte nach Cook, die nach den Kriterien für mittelgroße Stichproben auffällig sind. Eine weitere Möglichkeit besteht darin, auf robuste Regressionsverfahren zurückzugreifen, die wenig sensitiv auf Ausreißer und einflussreiche Werte reagieren. Solche Verfahren sind u. a. bei Wilcox (2012) beschrieben.

19.13.4 Multikollinearität

Unter Multikollinearität versteht man eine hohe multiple Korrelation eines Prädiktors mit anderen Prädiktoren. Wie wir anhand von Gleichung F 19.39 gesehen haben, wirkt sich eine hohe Multikollinearität dahingehend aus, dass der Standardfehler des Regressionsgewichts derjenigen Variablen, die mit den anderen hoch korreliert, groß ist und das Regressionsgewicht somit unpräzise geschätzt wird. Zur Bestimmung des Ausmaßes der Multikollinearität können zwei Koeffizienten bestimmt werden, die voneinander abhängen: der Toleranz- und der Varianzinflations-Faktor.

Toleranzfaktor und Varianzinflationsfaktor

Toleranzfaktor. Den Toleranzfaktor erhält man, indem man die quadrierte multiple Korrelation eines Prädiktors mit allen anderen unabhängigen Variablen von 1 abzieht:

$$TOL_j = 1 - R_{X_j|X_{\setminus j}}^2$$

Der Toleranzfaktor ist 0, wenn die Variable perfekt linear abhängig (perfekt vorhersagbar) von allen anderen Prädiktoren ist. Man spricht dann von exakter Multikollinearität. In diesem Fall kann die unabhängige Variable über die anderen Variablen hinaus nichts erklären, und ihr Regressionsgewicht kann auch mathematisch nicht eindeutig bestimmt werden. Der maximale Wert der Toleranz ist 1. In diesem Fall ist die Variable mit allen anderen Variablen unkorreliert. In der Literatur findet man häufig den Hinweis, dass ein Wert des Toleranzfaktors kleiner als 0,10 Multikollinearität anzeige, wobei auch bei größeren Werten Probleme auftreten können. In jedem Fall sollten die geschätzten Parameter und ihre Standardfehler bei jeder Anwendung genau inspiziert werden, und es sollte überprüft werden, ob die Regressionsgewichte plausible Werte annehmen und ihre Standardfehler nicht zu groß sind.

Varianzinflations-Faktor. Der Varianzinflations-Faktor ist der Kehrwert der Toleranz:

$$VIF_j = 1/TOL_j$$

Seine untere Grenze ist 1 und zeigt an, dass die Variable mit allen anderen unabhängigen Variablen unkorreliert ist. Je größer der Varianzinflations-Faktor wird, desto größer ist die Multikollinearität. Der Begriff Varianzinflations-Faktor lässt sich darauf zurückführen, dass die Varianz der Parameterschätzung, d.h. der quadrierte Standardfehler eines Regressionskoeffizienten, von der multiplen Korrelation beeinflusst wird. Ein Wert des Varianzinflations-Faktors, der größer als 10 ist, wird in der Literatur häufig als auffallend bewertet.

Beispiel

Zufriedenheit mit einem Kurs und wahrgenommene Anforderung: Toleranz- und Varianzinflations-Faktor

In unserem Lehrevaluationsbeispiel gibt es nur zwei unabhängige Variablen, die lineare und die quadratische Komponente. Es gibt daher nur eine Korrelation zwischen den unabhängigen Variablen, und folglich haben beide dieselben Werte auf dem Toleranz- und Varianzinflations-Faktor. Im Falle der unzentrierten Variablen beträgt der Toleranzfaktor $TOL_1 = TOL_2 = 0,046$ und der Varianzinflations-

Faktor $VIF_1 = VIF_2 = 21,739$. Es liegt also sehr hohe Multikollinearität vor. Im Falle der zentrierten Variablen liegen die Werte bei $TOL_1 = TOL_2 = 0,980$ und $VIF_1 = VIF_2 = 1,020$. Die Multikollinearität ist also äußerst gering.

Behebung des Multikollinearitätsproblems

Zur Behebung des Multikollinearitätsproblems gibt es verschiedene Möglichkeiten.

Zentrierung. Bei Produktvariablen, die sich aus einer geraden Anzahl von multiplizierten Variablen ergeben, wird das Multikollinearitätsproblem vermindert, indem man die Variablen zentriert. Dies gilt sowohl für die moderierte Regression als auch die nicht-lineare Regression mit polynomialen Termen.

Eliminierung von unabhängigen Variablen. Man kann unabhängige Variablen, die sehr hoch mit allen anderen unabhängigen Variablen korreliert sind, auch aus der Regressionsgleichung herausnehmen, sofern sie keinen eigenständigen Betrag über die anderen unabhängigen Variablen hinaus liefern.

Aggregation. Eine weitere Möglichkeit besteht darin, hoch interkorrelierte unabhängige Variablen zusammenzufassen (zu aggregieren). Der Aggregation liegt dabei die Annahme zugrunde, dass die hoch interkorrelierten Variablen aus psychologischer Sicht dasselbe Merkmal erfassen. Hierzu bildet man z. B. den Mittelwert aus den Variablen und nimmt die so neu gebildete aggregierte Variable anstelle der ursprünglichen Variablen in die Regressionsgleichung auf.

Faktorenanalytische Reduktion. Hoch interkorrelierte unabhängige Variablen können auch einer Faktorenanalyse oder einer Hauptkomponentenanalyse unterzogen werden (s. Kap. 25), und die Faktorwerte können als Werte der unabhängigen Variablen in die Regressionsanalyse aufgenommen werden. Schließlich besteht noch die Möglichkeit, die Abhängigkeit der Variablen dadurch zu berücksichtigen, dass für sie ein Faktor modelliert wird und dieser direkt in einem linearen Strukturgleichungsmodell als unabhängige Variable berücksichtigt wird (s. Kap. 26).

19.13.5 Homoskedastizität

Homoskedastizität bedeutet, dass sich die bedingten Residualvarianzen in der Population nicht voneinander unterscheiden. Die Varianz der Residuen hängt in der Population nicht von den Ausprägungen der unabhängigen Variablen ab und ist für die verschiedenen Konstellationen der unabhängigen Variablen gleich. Das Gegenteil von Homoskedastizität wird *Heteroskedastizität* genannt. Heteroskedastizität liegt vor, wenn die Varianzen der Residuen sich zwischen Ausprägungen der unabhängigen Variablen unterscheiden.

Konsequenzen der Heteroskedastizität

Ist die Homoskedastizität verletzt, so sind die geschätzten Parameter weiterhin erwartungstreue Schätzer für die Populationsparameter. Allerdings werden die Standardfehler und somit die Konfidenzintervalle und die Irrtumswahrscheinlichkeiten verzerrt bestimmt.

Residuenplots

Die Homoskedastizität lässt sich anhand von sog. Residuenplots darstellen, in denen üblicherweise studentisierte Residuen auf der y -Achse gegen die aufgrund der Regression vorhergesagten \hat{y} -Werte auf der x -Achse abgebildet werden. Abbildung 19.16 enthält einen Residuenplot für unser Lehrevaluationsbeispiel. Auf der x -Achse sind die korrigierten geschätzten Werte abgetragen. Das sind die geschätzten \hat{y} -Werte der Personen, die man anhand der Regressionskoeffizienten erhält, für deren Schätzung die Person aus der Stichprobe entnommen wurde. Auf der y -Achse sind die studentisierten Residualwerte abgetragen. Wie man sieht, schwanken

die Residualwerte unsystematisch um 0. Es lassen sich keine starken Unterschiede in den bedingten Schwankungen der Residualwerte erkennen. Die Annahme der Homoskedastizität kann beibehalten werden.

Abbildung 19.17 zeigt einen konstruierten Residuenplot, in dem die Homoskedastizitätsannahme verletzt ist. Die bedingten Varianzen sind an den Randbereichen stärker als im mittleren Bereich. Der Residuenplot zeigt noch eine weitere Besonderheit:

Die Residualwerte schwanken nicht unsystematisch um 0, sondern weisen eine kurvilineare Form auf. Dies zeigt, dass das Modell falsch spezifiziert wurde. Für den Residuenplot wurde ein Datensatz erzeugt, in dem eine kurvilineare Beziehung zwischen der abhängigen und der unabhängigen Variablen besteht, wobei in der Analyse nur die lineare Komponente spezifiziert wurde. Diese Fehlspezifikation kann durch den Residuenplot aufgedeckt werden. Der Residuenplot ist also nicht nur in der Lage, die Heteroskedastizität aufzudecken, sondern auch Fehlspezifikationen sichtbar zu machen. Fehlspezifikationen äußern sich darin, dass die bedingten Mittelwerte der Residuen von 0 abweichen. In unserem Beispiel sind die bedingten Mittelwerte der Residuen im mittleren Bereich negativ und in den Randbereichen positiv und weisen somit auf Fehlspezifikation hin.

Tests zur Überprüfung der Homoskedastizitätsannahme

Zur Überprüfung der Homoskedastizitätsannahme gibt es verschiedene statistische Tests, die bei Kockelkorn (2000) dargestellt sind. Diese Tests unterscheiden sich in den Alternativhypothesen, die sie testen, und setzen

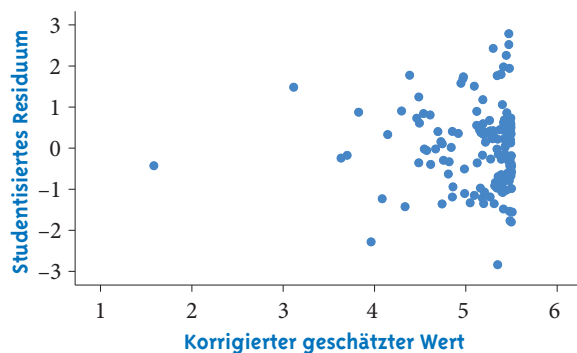


Abbildung 19.16 Residuenplot für das Beispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung«

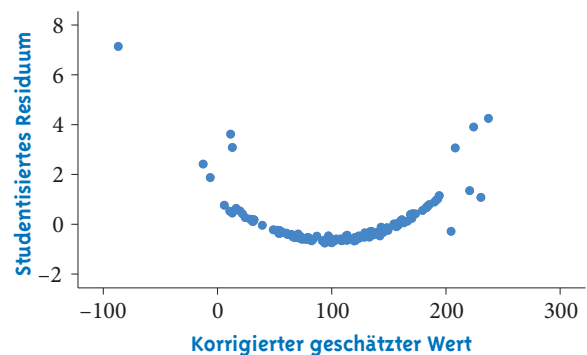


Abbildung 19.17 Residuenplot für das Beispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung«: Simulierte Daten

Vorwissen über die Form der Verletzung der Homoskedastizität voraus. Sie sind in Statistikprogrammen nicht routinemäßig implementiert und werden im Forschungsalltag selten angewendet. Als einfache Überprüfungsmöglichkeit schlagen Cohen et al. (2003) vor, die Untersuchungsobjekte anhand ihrer x -Werte in Gruppen annähernd gleicher Größe zu unterteilen und die Residualvarianzen in diesen Gruppen zu schätzen. Unterscheiden sich die beiden Gruppen mit der geringsten und größten Varianz um mindestens den Faktor 10, zeigt dies nach Cohen et al. (2003) eine bedeutende Verletzung der Homoskedastizitätsannahme an.

Umgang mit Heteroskedastizität

Im Falle von Heteroskedastizität kann auf das gewichtete Kleinst-Quadrate-Schätzverfahren (engl. weighted least squares; WLS) zurückgegriffen werden. Während beim einfachen Kleinst-Quadrate-Schätzverfahren (engl. ordinary least squares) die Regressionskoeffizienten so geschätzt werden, dass die Summe der quadrierten Abweichungen der beobachteten y -Werte von den vorhergesagten \hat{y} -Werten minimal ist, wird beim gewichteten Verfahren jeder quadrierte Abweichungswert mit einem individuellen Wert w_m gewichtet und die Summe der gewichteten quadrierten Abweichungen minimiert. Die Grundidee hierbei ist es, die y -Werte so zu transformieren, dass die Voraussetzung der Homoskedastizität erfüllt wird (Draper & Smith, 1998). Das Gewicht ist der Kehrwert der Residualvarianz in der Population für eine bestimmte Konstellation der Werte auf den unabhängigen Variablen, die für eine Person m mit einer entsprechenden Konfiguration der Werte auf der unabhängigen Variablen geschätzt wird. Um diese bedingte Residualvarianz zu erhalten, geht man in zwei Schritten vor: Im ersten Schritt schätzt man für jede Person ihren Residualwert in der Regressionsanalyse. In einem zweiten Schritt gehen die quadrierten Residualwerte als Werte der abhängigen Variablen in eine weitere Regression ein, in der sie auf die unabhängigen Variablen zurückgeführt werden. Die geschätzten Werte der abhängigen Variablen, deren Werte die quadrierten Residualwerte sind, sind dann die geschätzten Residualvarianzen für eine bestimmte Konstellation von unabhängigen Variablen. Die geschätzten Residualvarianzwerte müssen dann dem Datensatz hinzugefügt werden, und die Variable, die diese Werte enthält, muss als Gewichtsvariable definiert werden. Dieses Verfahren setzt allerdings große

Stichproben voraus, um sicherzustellen, dass es genügend Personen mit derselben Merkmalskonstellation bezüglich der unabhängigen Variablen gibt, um die bedingten Residualvarianzen zuverlässig schätzen zu können. Darüber hinaus weisen Cohen et al. (2003) darauf hin, dass in der WLS-Regressionsanalyse standardisierte Kenngrößen wie der Determinationskoeffizient keine klare Bedeutung mehr haben. Sie empfehlen daher, die WLS-Regression nur für den Fall sehr großer Stichproben und sehr gravierender Verletzungen der Homoskedastizität einzusetzen.

19.13.6 Unabhängigkeit der Residuen

In der Regressionsanalyse wird angenommen, dass die Residuen voneinander unabhängig sind. Diese Annahme ist in zwei typischen Anwendungsfällen in der Psychologie verletzt: (1) wenn der Stichprobenziehung Klumpenstichproben oder mehrstufige Auswahlverfahren zugrunde lagen (s. Abschn. 9.3.2), (2) bei serialer Abhängigkeit, die typischerweise in Einzelfalluntersuchungen auftritt (s. Abschn. 9.3.2).

Klumpenstichproben und mehrstufige Auswahlverfahren

Bezogen auf unser Beispiel der Lehrevaluation lägen Klumpenstichproben vor, wenn aus allen Lehrveranstaltungen Deutschlands eine Zufallsstichprobe gezogen worden wäre und man dann alle Studierenden der Lehrveranstaltung befragt hätte. Ein mehrstufiges Auswahlverfahren läge vor, wenn man zunächst aus den Dozenten Deutschlands eine Zufallsstichprobe gezogen, dann aus allen Lehrveranstaltungen des Dozenten eine Zufallsstichprobe ausgewählt und schließlich aus jeder Lehrveranstaltung eine Zufallsstichprobe von Studenten gezogen hätte. Diesen beiden Stichprobenziehungen ist gemeinsam, dass die Messobjekte (Studierende) keine einfache Zufallsstichprobe aus einer gemeinsamen Population darstellen, sondern Abhängigkeiten dadurch entstehen, dass sie gezielt aus einer gemeinsamen Lehrveranstaltung ausgewählt wurden. Die Annahme der Unabhängigkeit der Residuen wäre bei einer solchen Stichprobe z.B. dann verletzt, wenn sich Studierende, die dasselbe Anforderungsniveau wahrnehmen, in ihrer Zufriedenheitseinschätzung einander ähnlicher sind, wenn sie dieselbe Veranstaltung besucht haben, als wenn sie verschiedene Veranstaltungen besucht haben.

Konsequenzen. Liegt eine solche Abhängigkeit der Residuen vor, so werden die Regressionskoeffizienten nicht verzerrt geschätzt. Allerdings werden die Standardfehler typischerweise unterschätzt, sodass Effekte eher signifikant werden. Die faktische Irrtumswahrscheinlichkeit α ist also größer als das nominell festgelegte Signifikanzniveau.

Umgang mit dem Problem. Wie geht man mit dem Problem um, dass die Residuen voneinander abhängig sind? Eine Möglichkeit besteht darin, auf hierarchische lineare Modelle zurückzugreifen, die wir in Kapitel 20 behandeln werden. Diese Modelle berücksichtigen die geschachtelte Datenstruktur angemessen. Hat man nur eine geringe Anzahl von Klumpen oder Gruppen im mehrstufigen Auswahlverfahren, kann man auch einfach so vorgehen, dass man die Zugehörigkeit zu einem Klumpen oder einer Gruppe anhand von Codiervariablen codiert und diese Codiervariablen als unabhängige Variablen mit in die Regressionsgleichung aufnimmt (s. Abschn. 19.11).

Serielle Abhängigkeit

Dieses Problem liegt dann vor, wenn die Werte einer Ordnung unterliegen und die Residuen sich beeinflussen. Ein typischer Anwendungsbereich, in dem dieses Problem auftritt, ist die Einzelfallanalyse. Angenommen, wir befragen eine Person jeden Morgen nach ihrer Stimmung und erheben, wie lange die Person geschlafen hat. Uns interessiert die regressive Abhängigkeit der Stimmung von der Schlafdauer. Die Schlafdauer wird die Stimmung nicht perfekt vorhersagen. Da die Stimmung von vielen Einflüssen abhängt, wird es Tage geben, an denen eine Person bei gleicher Schlafdauer besser oder schlechter gestimmt ist, als man allein aufgrund der Schlafdauer erwarten würde. Darüber hinaus kann die Stimmung von der Stimmung am Vortag abhängen. Geht man schlecht gelaunt ins Bett, kann die Stimmung am nächsten Morgen schlechter sein, als man aufgrund der Schlafdauer erwarten würde. Schläft man wohlgemut ein, wird die Stimmung am nächsten Tag bei gleicher Schlafdauer tendenziell besser sein. Dies hat zur Folge, dass das Residuum der Messung an einem Tag vom Residuum des Vortages abhängt. War das Residuum am Vortag positiv (bessere Stimmung als aufgrund der Schlafdauer zu erwarten), wird tendenziell das Residuum am Folgetag auch positiver sein. Die Residuen der beiden Tage hängen somit voneinander

ab. Sofern es sich um einen systematischen Prozess handelt, kann man die sog. Autokorrelationen der Residuen bestimmen.

Autokorrelation. Eine Autokorrelation ist die Korrelation eines Merkmals mit seiner zeitversetzt wiederholten Messung. Man unterscheidet bei Autokorrelationen die Ordnung der Autokorrelation. Bei einer Autokorrelation erster Ordnung wird eine Messreihe mit der um einen Messzeitpunkt versetzten Messreihe korreliert. Bei einer Autokorrelation zweiter Ordnung ist die Wiederholungsmessreihe um zwei Messzeitpunkte versetzt etc. Eine Autokorrelation erster Ordnung der Residuen zeigt an, wie stark die Residuen des Folgetages mit den Residuen des Vortages (über die gesamte Messreihe hinweg) zusammenhängen. Liegen solche Autokorrelationen vor, ist die Annahme der Unabhängigkeit der Residuen verletzt. Die Regressionsgewichte werden zwar unverzerrt geschätzt, aber die Standardfehler sind nicht korrekt. In einem solchen Fall sollte auf regressionsanalytische Modelle, die diese Abhängigkeit berücksichtigen, zurückgegriffen werden. Für den Einzelfall sind z. B. zeitreihenanalytische Ansätze die Modelle der Wahl (z. B. Köhler, 2008). Ob die Autokorrelation der Residuen statistisch bedeutsam ist, kann mit dem Durbin-Watson-Test überprüft werden, der in vielen Statistikprogrammen implementiert ist.

19.13.7 Normalverteilung der Residuen

Im Allgemeinen Linearen Modell wird angenommen, dass die Residualvariablen in der Population bedingt – und daher auch unbedingt – normalverteilt sind und sich deren bedingte Varianzen nicht unterscheiden. Allerdings haben wir schon in Abschnitt 19.13.3 gesehen, dass die geschätzten Residualwerte nicht die gleichen bedingten Varianzen aufweisen. Im Gegenteil, sie unterscheiden sich in ihnen. Wir haben dies bereits bei der Bestimmung der studentisierten Residuen behandelt. Diese erhält man, indem man die Residualwerte durch die bedingten geschätzten Standardabweichungen dividiert. Auch sind die geschätzten Residualwerte in systematischer Weise voneinander abhängig (Kockelkorn, 2000). Daher werden zur Überprüfung der Normalverteilungsannahme keine exakten Tests herangezogen, sondern heuristische Verfahren. Zum einen kann man

sich das Histogramm der Residuen anschauen, zum anderen einen Probability-Probability-Plot.

Histogramm der Residuen

Kockelkorn (2000) empfiehlt, ein Histogramm der studentisierten Residuen zu erstellen, die er skalierte Residuen nennt, da diese identisch normalverteilt sind. Allerdings erstellen Statistikprogramme wie SPSS routinemäßig bei der Regressionsdiagnostik häufig nur das Histogramm der standardisierten Residuen, sodass das Histogramm der studentisierten Residuen vom Anwender erstellt werden muss (bei SPSS mit der Option »deskriptive Statistik«). In Abbildung 19.18 ist das Histogramm der studentisierten Residuen für das Lehrevaluationsbeispiel dargestellt. Es gibt keinen Anlass, die Hypothese der Normalverteilung der Residuen in der Population zu verwerfen. In der Tat haben wir die Daten aus didaktischen Gründen so simuliert, dass dem Modell der Datensimulation eine normalverteilte Residualvariable zugrunde lag. Wie Abbildung 19.18 verdeutlicht, kann man allein schon aufgrund des Stichprobenfehlers nicht davon ausgehen, dass die geschätzten Residualwerte perfekt einer Normalverteilung folgen.

Probability-Probability-Plot

Eine zweite Möglichkeit besteht darin, den sog. Probability-Probability-Plot (P-P-Plot) zu inspizieren, in dem zwei kumulierte Wahrscheinlichkeiten gegeneinander abgetragen werden. Auch für diesen Plot sollte man auf die studentisierten Residuen zurückgreifen, wohingegen Statistikprogramme wie SPSS häufig routinemäßig

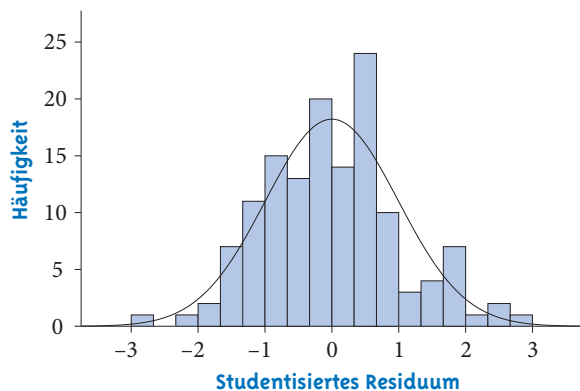


Abbildung 19.18 Histogramm der studentisierten Residuen für das Beispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung«

die standardisierten Residuen zugrunde legen. Auch der P-P-Plot setzt eigentlich voraus, dass die Beobachtungen voneinander unabhängig sind, was bei den geschätzten Residuen nicht der Fall ist. Nichtsdestotrotz ist der P-P-Plot von großem heuristischem Wert. In Abbildung 19.19 ist ein solcher P-P-Plot der studentisierten Residuen für unser Beispiel dargestellt, anhand dessen man die Normalverteilungshypothese visuell überprüfen kann. Auf der Abszisse sind die geschätzten kumulierten Wahrscheinlichkeiten der studentisierten Residuen angegeben. Hierzu ordnet man die Residuen ihrer Größe nach und schätzt dann die kumulierten Wahrscheinlichkeiten anhand der kumulierten Häufigkeiten der Daten. Auf der Ordinate sind dazu die kumulierten Wahrscheinlichkeiten für einen spezifischen Abszissenwert aufgetragen, die man bei Gültigkeit eines bestimmten Verteilungsmodells erwarten würde. In unserem Fall wollen wir überprüfen, ob die Werte normalverteilt sind. Die Normalverteilung hängt von zwei Parametern ab, dem Erwartungswert und der Varianz. Da man beide Populationsparameter nicht kennt, schätzt man sie aus den Daten. Daraufhin schätzt man die erwarteten kumulierten Wahrscheinlichkeiten über die Quantile der Normalverteilung. Sind die Residuen

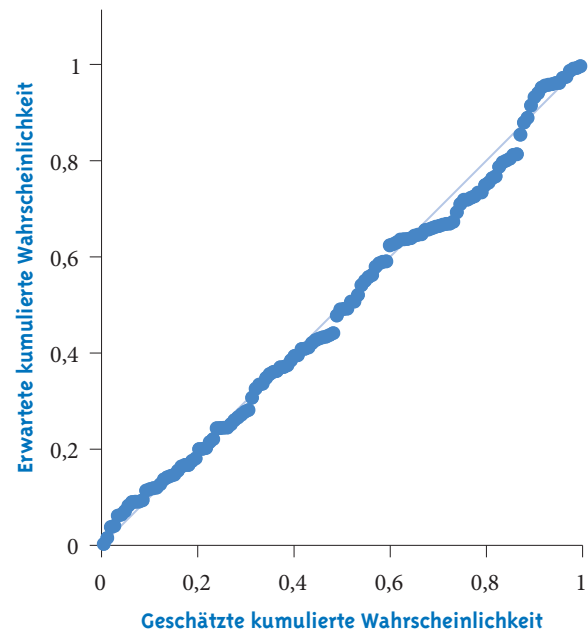


Abbildung 19.19 Probability-Probability-Plot der studentisierten Residuen für das Beispiel »Zufriedenheit mit einem Kurs und wahrgenommene Anforderung«

normalverteilt, liegen alle Punkte auf einer Geraden. In unserem Fall weichen sie nur unwesentlich von der Geraden ab. Auch dies zeigt, dass die Normalverteilungsannahme nicht gravierend verletzt ist.

Neben dem P-P-Plot gibt es noch den Q-Q-Plot (Quantile-Quantile-Plot), der dieselbe Information enthält. Beim Q-Q-Plot werden nicht die kumulierten Wahrscheinlichkeiten, sondern die Quantile gegeneinander abgetragen. Auch beim Q-Q-Plot liegen im Idealfall alle Punkte auf einer Geraden.

Konsequenzen der Verletzung der Normalverteilungsannahme

Ist die Normalverteilungsannahme verletzt, erhält man trotzdem unverzerrte Schätzungen der Regressionsgewichte. Bei großen Stichproben wirkt sich die Verletzung auch nicht gravierend auf die Schätzung der Standardfehler und die Signifikanztests aus. Hinweise auf die Verletzung der Normalverteilungsannahme, die man durch die Histogramme der studentisierten Residuen und den P-P-Plot erhält, können aber auf eine Fehlspezifikation des Modells hinweisen (Cohen et al., 2003). Daher werden beide grafischen Methoden v. a. dazu genutzt, fehlspezifizierte Modelle aufzudecken. Um dies zu verdeutlichen, haben wir Daten derart simuliert, dass es eine starke kurvilineare Abhängigkeit zwischen der abhängigen und einer unabhängigen Variablen gibt. In die Regressionsanalyse haben wir allerdings nur die lineare Komponente aufgenommen. Das Histogramm und der P-P-Plot der studentisierten Residuen in den Abbildungen 19.20 und 19.21 zeigen eine starke Abweichung von der Normalverteilungsannahme an. Diese ist aber ausschließlich auf die Fehlspezifikation des Modells zurückzuführen. Nimmt man die quadratische Komponente ins Modell auf, weisen die Residuen keine gravierende Abweichung von der Normalverteilung mehr auf. Dies verwundert nicht, da der Datensimulation normalverteilte Residuen zugrunde lagen.

Umgang mit nicht-normalverteilten Daten

Zeigen die Residuen Abweichungen von der Normalverteilung an, ist zunächst zu überprüfen, ob dies auf Fehlspezifikationen zurückgeführt werden kann. Ist dies nicht der Fall und hat man kleine Stichproben, sodass sich die Verletzung der Normalverteilungsannahme auf die Standardfehler, Konfidenzintervalle und Signifikanztests auswirken kann, können die Daten so

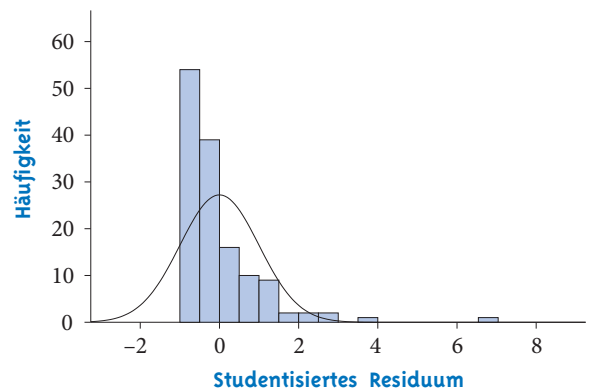


Abbildung 19.20 Histogramm der studentisierten Residuen: Simulierte Daten

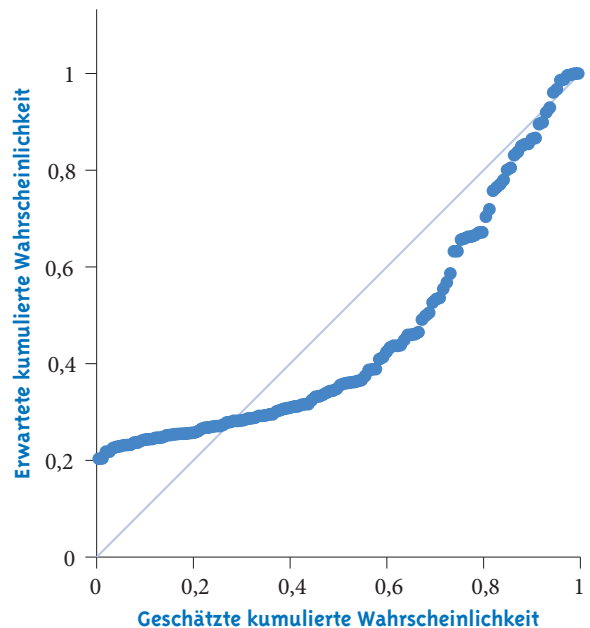


Abbildung 19.21 Probability-Probability-Plot der studentisierten Residuen: Simulierte Daten

transformiert werden, dass ihre Abweichung von der Normalverteilung geringer wird. Hierzu kann je nach Verteilung der Variablen auf verschiedene Transformationen zurückgegriffen werden. Wir wollen nur zwei Transformationen illustrieren. Die Verteilung der Variablen in Abbildung 19.22 a zeigt eine Verteilungsform, die der Exponentialverteilung ähnelt. Folglich transformieren wir die Daten mit einer logarithmischen Transformation der Art $X_t = \ln(X)$, wobei X_t die transformierte Variable bezeichnet und »ln« für den Logarithmus naturalis steht. Die Verteilung der trans-

formierten Variablen finden wir in Abbildung 19.22 c. Die transformierte Variable folgt approximativ einer Normalverteilung. Die Verteilung in Abbildung 19.22 b ist extrem rechtssteil und linksschief. Diese Variable transformieren wir mit der Transformationsvorschrift

$$X_t = \frac{1}{2} \cdot \ln\left(\frac{1+X}{1-X}\right)$$

und erhalten wiederum die Verteilung in Abbildung 19.22 c, die approximativ normalverteilt ist.

Um eine optimale Transformation der Variablen auf die Normalverteilung zu finden, kann man sich auch Transformationsverfahren wie die Box-Cox-Transformation (Box & Cox, 1964) zunutze machen, die es erlaubt aus einer Familie von Transformationen diejenige zu finden, die die Variable so transformiert, dass sich ihre Verteilung der Normalverteilung maximal annähert (s. zur Einführung z. B. Osborne, 2010).

Andere Regressionsmodelle

Die Voraussetzung der Normalverteilung macht deutlich, dass der multiplen Regressionsanalyse stetige abhängige Variablen zugrunde liegen. Nicht in allen Anwendungskontexten hat man jedoch stetige oder quasi-stetige abhängige Variablen vorliegen, sondern andere Variablenarten, bei denen die Normalverteilungsannahme a priori nicht sinnvoll ist. Es wurden daher auch Regressionsmodelle für andere Arten von abhängigen Variablen entwickelt.

Kategoriale Variablen. Sind die abhängigen Variablen kategorial, so greift man auf die logistische Regressionsanalyse zurück. Wir werden in Kapitel 22 die logistische Regressionsanalyse für nominalskalierte und ordinalskalierte abhängige Variablen behandeln.

Häufigkeitsvariablen. Sind die Werte der abhängigen Variablen Häufigkeiten (z. B. Anzahl der Facebook-Postings in einer Woche, Anzahl der im Leben erlebten Fahrradunfälle, Anzahl der gelösten Testaufgaben etc.), kann man auf die Poisson-Regression zurückgreifen. Hervorzuheben ist, dass es sich bei den Werten einer Häufigkeitsvariablen um individuelle Häufigkeiten handelt (z. B., wie häufig jemand in einer Stunde seine E-Mails checkt). Die Untersuchungseinheiten unterscheiden sich in diesen individuellen Häufigkeiten. Die Poisson-Regression ist der Regression, so wie wir sie in diesem Kapitel behandeln, insbesondere dann überlegen, wenn der Mittel-

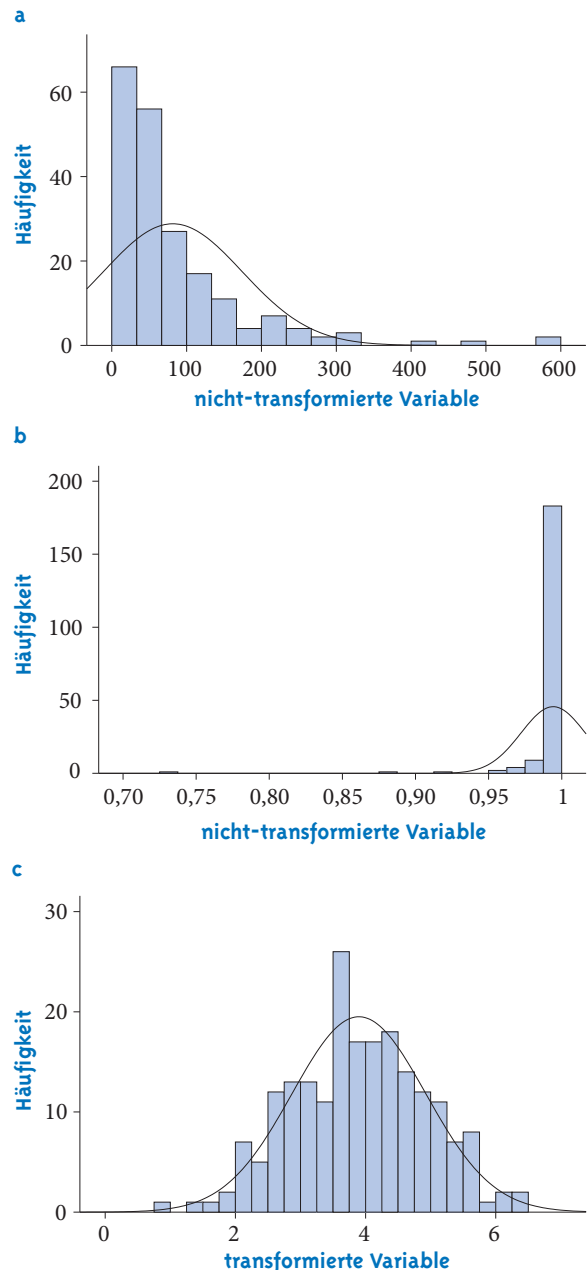


Abbildung 19.22 Zwei Typen nicht-normalverteilter Variablen und ihre Verteilungsform nach ihrer Transformation

wert der Häufigkeitsvariablen klein (< 10) ist (Coxe et al., 2009). Coxe et al. (2009) geben eine Einführung in die verschiedenen Varianten der Poisson-Regression und zeigen, anhand welcher Kriterien man für eine konkrete Fragestellung die beste Variante auswählt und was die Parameterschätzungen bedeuten.

Skalen mit beschränkten Wertebereich. Skalen, die in den Verhaltenswissenschaften eingesetzt werden, weisen häufig einen beschränkten Wertebereich auf. Hierzu gehören visuelle Analogskalen (Eid & Schmidt, 2014). Ein Beispiel ist die Einschätzung des Schmerzempfindens auf einer Linie, die 10 cm lang ist und deren Endpunkte mit den Zahlen 0 und 100 verankert sind. Ein anderes Beispiel für Skalen mit beschränktem Wertebereich ist die Einschätzung der Wahrscheinlichkeit, an der nächsten Bundestagswahl teilzunehmen, oder die Einschätzung der Wahrscheinlichkeit, an Krebs zu erkranken. Solche Skalen weisen häufig Verteilungen auf, die J- oder U-förmig sind. Arostegui et al. (2010), Kieschnik und McCullough (2003) sowie Lesaffre et al. (2007) geben einen Überblick über regressionsanalytische Verfahren, die bei solchen abhängigen Variablen eingesetzt werden können.

19.13.8 Multivariate Normalverteilung der Variablen

Die Verteilungsannahmen in Bezug auf die Residuen werden im Modell mit deterministischen Regressoren getroffen. In diesem Modell müssen keine Verteilungsannahmen in Bezug auf die unabhängigen Variablen getroffen werden. Im Modell mit stochastischen Regressoren wird eine Verteilungsannahme in Bezug auf die gemeinsame Verteilung der abhängigen und aller unabhängigen Variablen getroffen. Obwohl auch andere Verteilungsannahmen denkbar sind und entsprechende Ansätze formuliert wurden, wird im Modell mit stochastischen Regressoren typischerweise die Annahme der multivariaten Normalverteilung getroffen. Wie diese Annahme überprüft werden kann, haben wir schon im Abschnitt 16.4.1 zur Produkt-Moment-Korrelation behandelt.

Tests auf multivariate Normalverteilung finden aus dreierlei Gründen selten Anwendung in der Forschungspraxis:

- (1) Sie sind in vielen Standard-Statistikprogrammen nicht implementiert.
- (2) Aus der Verletzung der univariaten Normalverteilung, für deren Überprüfung verschiedene Standardtests zur Verfügung stehen (s. Abschn. 10.6), folgt auch die Verletzung der multivariaten Normalverteilung (allerdings folgt aus der Nichtverletzung der univariaten Normalverteilung nicht

zwangsläufig die Nichtverletzung der multivariaten Normalverteilung).

- (3) Bei vielen Anwendungen geht man implizit vom Modell mit deterministischen Regressoren aus.

Auch wenn dieses Modell für viele Fragestellungen der Psychologie und anderer Sozial- und Verhaltenswissenschaften streng genommen nicht das korrekte Modell ist, schränkt man – um die Vorteile und Einfachheit des Allgemeinen Linearen Modells zu nutzen – implizit die Ergebnisse auf die realisierten Werte der unabhängigen Variablen ein bzw. geht davon aus, sie in zukünftigen Studien wiederherstellen zu können. Oder aber man geht wie O'Brien und Muller (1993) davon aus, dass sich beide Modelle in ihren Ergebnissen in Bezug auf das Konfidenzintervall des Determinationskoeffizienten und die Bestimmung der optimalen Stichprobengröße nicht gravierend unterscheiden. Wie wir in unserem Beispiel in Abschnitt 19.7 gesehen haben, war dieser Unterschied im konkreten Anwendungsfall nicht groß.

19.13.9 Verletzung der Annahmen und Konsequenzen

In Tabelle 19.29 ist zusammengestellt, wie sich Verletzungen von Annahmen auf die Schätzung der Regressionsgewichte und der Standardfehler auswirken.

Tabelle 19.29 Konsequenzen der Verletzung von Voraussetzungen der multiplen Regressionsanalyse

Voraussetzung	Geschätzte Regressionskoeffizienten	Geschätzte Standardfehler
Korrekte Spezifikation des Modells	Verzerrung	Verzerrung
Messfehlerfreiheit der unabhängigen Variablen	Verzerrung	Verzerrung
Homoskedastizität	keine Verzerrung	Verzerrung
Unabhängigkeit der Residuen	keine Verzerrung	Verzerrung
Normalverteilung der Residuen	keine Verzerrung	Verzerrung

Es zeigt sich, dass sich nur eine Fehlspezifikation des Modells und Messfehler bei den unabhängigen Variablen dahin gehend auswirken, dass die Regressionsparameter verzerrt geschätzt werden. Verletzungen aller Voraussetzungen wirken sich hingegen auf verzerrte Schätzungen der Standardfehler aus und haben somit verzerrte Signifikanztests zur Folge. In der Tabelle erscheinen Ausreißer und einflussreiche Daten sowie die Multikollinearität nicht, da es sich hierbei nicht um Voraussetzungen der multiplen Regressionsanalyse handelt, sondern um problematische Datensituationen.

Zusammenfassung

- ▶ Die multiple Regressionsanalyse dient der Kontrolle von Störvariablen, der Prognose und Erklärung des Verhaltens anhand mehrerer unabhängiger Variablen.
- ▶ Die multiple Regressionsanalyse ist ein kompensatorisches Modell – niedrige Werte auf einer unabhängigen Variablen können durch hohe Werte auf anderen unabhängigen Variablen ausgeglichen werden.
- ▶ Die Regressionskoeffizienten werden nach dem Kleinste-Quadrate-Kriterium geschätzt.
- ▶ Das Regressionsgewicht einer unabhängigen Variablen gibt an, um wie viel sich die erwarteten \hat{y} -Werte ändern, wenn man die unabhängige Variable um eine Maßeinheit erhöht und die Werte auf allen anderen unabhängigen Variablen konstant hält.
- ▶ Regressionsgewichte in der multiplen Regression werden auch Partialregressionsgewichte genannt, da sie als einfache Regressionsgewichte von Residualvariablen dargestellt werden können.
- ▶ Auf unstandardisierte Regressionsgewichte greift man beim Vergleich mehrerer Gruppen zurück; standardisierte Regressionsgewichte eignen sich zum Vergleich verschiedener Variablen.
- ▶ Die multiple Korrelation ist die bivariate Korrelation zwischen der abhängigen Variablen und der aufgrund der unabhängigen Variablen vorhergesagten abhängigen Variablen. Sie kann Werte von 0 bis 1 annehmen. Ihr quadriert Wert ist der multiple Determinationskoeffizient.
- ▶ Der multiple Determinationskoeffizient kann in eine Summe von Semipartialdeterminationen zunehmend höherer Ordnung zerlegt werden.
- ▶ Der Anteil der Varianz der abhängigen Variablen, der von einer unabhängigen Variablen über alle anderen unabhängigen Variablen hinaus determiniert wird, heißt Nützlichkeit.
- ▶ Im Modell mit deterministischen Regressoren geht man davon aus, dass die Ausprägungen der unabhängigen Variablen feste Werte sind, die unter Kontrolle des Forschenden stehen. Zur inferenzstatistischen Absicherung der Modellgrößen wird angenommen, dass die Residualvariablen bedingt (d.h. gegeben die Ausprägungen der unabhängigen Variablen) normalverteilt sind und ihre bedingte Varianz für alle Konstellationen der unabhängigen Variablen konstant ist (Homoskedastizität). Darüber hinaus müssen die Residuen voneinander unabhängig sein.
- ▶ Im Modell mit stochastischen Regressoren wird angenommen, dass die Werte der unabhängigen Variablen Realisierungen von Zufallsvariablen sind und sich ihre Werte daher von Stichprobe zu Stichprobe unterscheiden können. In diesem Modell wird angenommen, dass die abhängigen und unabhängigen Variablen multivariat normalverteilt sind. Hieraus folgt, dass die Residuen bedingt normalverteilt sind und gleiche bedingte Varianzen aufweisen.
- ▶ Das Modell mit stochastischen Regressoren unterscheidet sich vom Modell mit deterministischen Regressoren nicht in den Parameterschätzungen, wohl aber im Konfidenzintervall für den Determinationskoeffizienten und in der Bestimmung der optimalen Stichprobengröße, um einen a priori festgelegten Populationsdeterminationskoeffizienten abzusichern.
- ▶ Die Effektgröße zur Bestimmung der optimalen Stichprobengröße besteht aus dem Populations-Determinationskoeffizienten dividiert durch den Populations-Indeterminationskoeffizienten.
- ▶ Datengesteuerte Verfahren zur Auswahl bedeutsamer unabhängiger Variablen sind die Vorwärtsselektion, die Rückwärtselimination und die schrittweise Regression.
- ▶ Die Prognose- und Kreuzvalidierungsfehler erlauben es abzuschätzen, wie zuverlässig die Regressionsanalyse in zukünftigen Studien eingesetzt werden kann.
- ▶ Suppressorvariablen lassen sich in klassische, reziproke und negative Suppressorvariablen unterteilen.
- ▶ In der moderierten Regressionsanalyse werden Interaktionen zwischen den unabhängigen Variablen berücksichtigt.

- ▶ In der moderierten Regressionsanalyse können die unabhängigen Variablen vor der Bildung der Produktvariablen zentriert werden, um die Interpretation der Regressionskoeffizienten zu erleichtern.
- ▶ Nicht-lineare Beziehungen lassen sich durch die Hinzunahme von Polynomen höherer Ordnung berücksichtigen.
- ▶ Kategoriale Variablen müssen durch Codiervariablen codiert werden, um sie als unabhängige Variablen in die Regressionsanalyse aufzunehmen.
- ▶ Einfache Codiermöglichkeiten sind die Dummy- und die Effektcodierung.
- ▶ Bei der Kovarianzanalyse werden metrische und kategoriale Variablen als unabhängige Variablen berücksichtigt. Es wird angenommen, dass sie additiv, nicht aber multiplikativ zusammenwirken.
- ▶ Adjustierte Mittelwerte sind die mittels der Kovarianzanalyse vorhergesagten Gruppenmittelwerte an der Stelle der Gesamtmittelwerte der Kovariaten.
- ▶ Wertet man Untersuchungspläne der quasi- und nicht-experimentellen Forschung mit der Kovarianzanalyse aus, müssen Artefakte, die durch den Messfehler und ausgelassene Drittvariablen hervorgerufen werden können, besonders beachtet werden.
- ▶ Bei der Aptitude-Treatment-Interaction-Analyse werden Interaktionen zwischen metrischen und kategorialen unabhängigen Variablen zugelassen.
- ▶ Die Regressionsdiagnostik widmet sich der Aufdeckung von Umständen, die die valide Interpretation der Ergebnisse gefährden.
- ▶ Um die Ergebnisse der Regressionsanalyse angemessen interpretieren zu können, ist es notwendig, dass das Modell korrekt spezifiziert wird. Fehlspezifikationen führen zu verzerrten Parameterschätzungen und verzerrten Standardfehlern.
- ▶ Bei der Regressionsanalyse wird angenommen, dass die unabhängigen Variablen messfehlerfrei gemessen wurden. Der Messfehler führt zu verzerrten Schätzungen der Regressionsparameter und ihrer Standardfehler.
- ▶ Verletzungen der Homoskedastizität können mit einem Residuenplot aufgedeckt werden. Heteroskedastizität führt zu verzerrten Standardfehlern.
- ▶ Die bedingte Normalverteilungsannahme der Residuen kann mit einem Histogramm und einem Probability-Probability-Plot der studentisierten Residuen grafisch überprüft werden. Verletzungen der Normalverteilungsannahme führen bei kleinen Stichproben zu verzerrten Schätzungen der Standardfehler.
- ▶ Die Abhängigkeit der Residuen kann u. a. durch einen Klumpeneffekt, mehrstufige Auswahlverfahren und seriale Abhängigkeiten hervorgerufen werden. Abhängigkeiten der Residuen führen zu verzerrten Schätzungen der Standardfehler.
- ▶ Multikollinearität kann durch den Toleranz- und den Varianzinflations-Faktor aufgedeckt werden. Hohe Multikollinearität führt zu hohen Standardfehlern der Regressionsgewichte.

Fragen und Übungsaufgaben

Fragen

- (1) Nennen Sie verschiedene Zielsetzungen, die man mit der Anwendung der multiplen Regressionsanalyse verfolgt.
- (2) Wie lautet die Regressionsgleichung für Merkmalsträger?
- (3) Warum handelt es sich bei der multiplen Regressionsanalyse um ein kompensatorisches Modell?
- (4) Unter welchen Bedingungen ist das Regressionsgewicht der multiplen Regressionsanalyse
 - (a) gleich dem
 - (b) kleiner als das
 - (c) größer als das
 Regressionsgewicht der unabhängigen Variablen in der einfachen Regressionsanalyse?
- (5) Wie hängt das Partialregressionsgewicht mit der Partialkorrelation zusammen?
- (6) Für welche Fragestellungen verwendet man unstandardisierte, für welche standardisierte Regressionsgewichte?
- (7) Was bedeutet die multiple Korrelation, und wie berechnet man sie?
- (8) Wie lautet die Quadratsummenzerlegung der multiplen Regressionsanalyse?
- (9) Was versteht man darunter, dass man die multiple Determination als Summe von Semipartialdeterminationen zunehmend höherer Ordnung darstellen kann?

- (10) Wie lautet das Populationsmodell der multiplen Regressionsanalyse?
- (11) Welche Voraussetzungen zur inferenzstatistischen Testung werden im Modell mit deterministischen Regressoren gemacht?
- (12) Welche Voraussetzungen zur inferenzstatistischen Testung werden im Modell mit stochastischen Regressoren gemacht?
- (13) Wie überprüft man inferenzstatistisch die Nullhypothese, dass
 - (a) der Determinationskoeffizient in der Population
 - (b) der quadrierte Semipartialkorrelationskoeffizient in der Population gleich 0 ist?
- (14) Wie lautet die Effektgröße ϕ_1^2 zur Bestimmung der optimalen Stichprobengröße?
- (15) Welche Verfahren zur Auswahl unabhängiger Variablen kennen Sie? Erläutern Sie diese.
- (16) Was versteht man unter dem Prognosefehler?
- (17) Was versteht man unter dem Kreuzvalidierungsfehler?
- (18) Was versteht man unter einer Suppressorvariablen?
- (19) Welche Typen von Suppressorvariablen unterscheidet man, und was bedeuten sie?
- (20) Was versteht man unter der Nützlichkeit einer Variablen?
- (21) Was versteht man unter Zentrierung von Variablen, und weswegen nimmt man sie vor?
- (22) Beschreiben Sie das Vorgehen zur Überprüfung von Interaktionseffekten in der multiplen Regressionsanalyse (moderierte Regression).
- (23) Was sind bedingte Regressionsgewichte?
- (24) Wie kann man mit der Regressionsanalyse nicht-lineare Zusammenhänge untersuchen?
- (25) Was versteht man unter Dummy-Codierung, und nach welchem Prinzip werden die Codiervariablen gebildet?
- (26) Worin unterscheiden sich die ungewichtete und die gewichtete Effektcodierung?
- (27) Was bedeuten die Regressionskoeffizienten bei der Dummy-, was bei der Effektcodierung?
- (28) Wie überprüft man Interaktionen zwischen kategorialen unabhängigen Variablen mit der multiplen Regressionsanalyse?
- (29) Was ist eine Kovarianzanalyse?
- (30) Was versteht man unter adjustierten Mittelwerten, und warum bestimmt man sie?
- (31) Welche Probleme stellen sich, wenn man die Kovarianzanalyse zur Auswertung quasi-experimenteller Untersuchungspläne in der Evaluationsforschung einsetzt?
- (32) Was ist Lords Paradox?
- (33) Was versteht man unter einer Aptitude-Treatment-Interaction-Analyse?
- (34) Was versteht man unter Under- und was unter Overfitting, und welche Konsequenzen haben diese?
- (35) Was ist ein LOWESS-Anpassungsverfahren?
- (36) Wie kann man Ausreißerwerte auf der abhängigen und der unabhängigen Variablen identifizieren?
- (37) Was sind einflussreiche Datenpunkte, und wie kann man sie identifizieren?
- (38) Was versteht man unter Multikollinearität, und wie kann man sie aufdecken?
- (39) Was sind die Konsequenzen von Heteroskedastizität?
- (40) Was ist ein Residuenplot, und wofür setzt man ihn ein?
- (41) Was versteht man unter der Unabhängigkeit von Residuen?
- (42) Wie kann man die Normalverteilungshypothese untersuchen?
- (43) Welche Konsequenzen haben die Verletzungen von Annahmen, auf denen die multiple Regressionsanalyse basiert?

Übungsaufgaben

- (1) Zeigen Sie, dass zwischen dem Partialregressionsgewicht und der Partialkorrelation die Beziehung F 19.13 gilt.
- (2) Zeigen Sie, dass Gleichung F 19.19 aus Gleichung F 19.18 folgt.
- (3) In einer Untersuchung zur Lebenszufriedenheit (Y) wollen Sie die Hypothese überprüfen, dass sich die Arbeitszufriedenheit (X_1) umso stärker auf die Lebenszufriedenheit auswirkt, je größer die Wichtigkeit ist, die man der Arbeit zuschreibt (X_2). Hierzu erheben Sie die Lebenszufriedenheit,

die Arbeitszufriedenheit und die Wichtigkeit mit intervallskalierten Skalen. Beschreiben Sie, wie Sie zur Überprüfung dieser Hypothese vorgehen. Formulieren Sie hierbei (mathematisch) das Regressionsmodell, das Sie Ihrer Hypothesenüberprüfung zugrunde legen, und formulieren Sie die statistische Nullhypothese, die Sie testen wollen. Beschreiben Sie auch, wie Sie zu einer statistischen Entscheidung gelangen.

- (4) In einer multiplen Regressionsanalyse wird der Einfluss der unabhängigen Variablen A, B, C und D auf die abhängige Variable E untersucht. Die einzelnen Variablen erfassen folgende Konstrukte:

A: Zufriedenheit mit dem Freundeskreis

B: Zufriedenheit mit dem Partner

C: Zufriedenheit mit der finanziellen Unterstützung durch die Eltern

D: Zufriedenheit mit den Wohnverhältnissen

E: Lebenszufriedenheit

Die einzelnen unabhängigen Variablen wurden nacheinander in die Regressionsanalyse aufgenommen. In der folgenden Tabelle sind die Werte der zugehörigen Determinationskoeffizienten angegeben:

Prädiktoren in der Regressionsgleichung	R^2
A	0,16
A, B	0,27
A, B, C	0,49
A, B, C, D	0,65

Ergänzen Sie die folgenden Aussagen, so dass sie richtige Antworten ergeben.

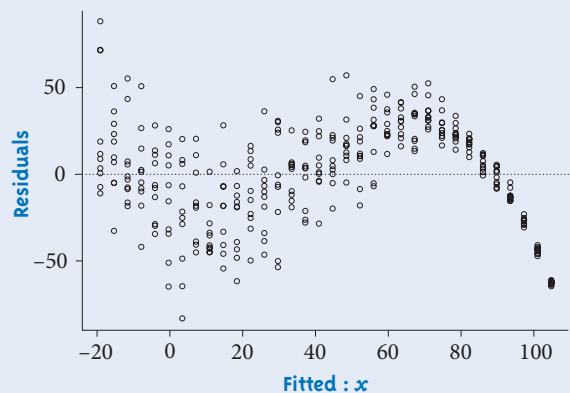
Die Produkt-Moment-Korrelation zwischen der Variablen E und der Variablen _____ beträgt $r = 0,40$.

Die _____ Korrelation zwischen der Variablen E und D beträgt 0,40.

Die multiple Korrelation zwischen der Variablen E und den Variablen A, B und C beträgt _____.

Die Variablen C und D erklären 38 % der Variation in der Lebenszufriedenheit unter der Bedingung, dass _____.

- (5) Nachdem Sie eine einfache lineare Regression gerechnet haben, finden Sie folgenden Residuenplot. Die Achse »Fitted: x« kennzeichnet die vorhergesagten Werte. Welche Schlussfolgerungen ziehen Sie in Bezug auf die Gültigkeit der Annahmen der Regressionsanalyse in diesem Anwendungsfall? Welche Annahmen werden verletzt sein? Warum?



- (6) In der folgenden Tabelle finden Sie zwei nominalskalierte Variablen. Sie möchten beide Variablen in eine Regressionsanalyse als unabhängige Variablen aufnehmen und neben ihren Haupteffekten auch den Effekt der Interaktion testen. Tragen Sie in die folgende Tabelle die Dummy-Variablen ein, um die Information, die in den unabhängigen Variablen steckt, zu codieren. Wählen Sie für jede Dummy-Variable eine Spalte, und geben Sie in den Zeilen für jede Dummy-Variable den Wert an, den die Dummy-Variable für diese Gruppe (Wertkombinationen) annimmt.

Variable 1 (Altersgruppe)	Variable 2 (Bildungsniveau)	Dummy-Variable
1	1	
2	1	
3	1	
1	2	
2	2	
3	2	

