

## Fragen und Antworten zu Kapitel 19

- (1) Nennen Sie verschiedene Zielsetzungen, die man mit der Anwendung der multiplen Regressionsanalyse verfolgt.**

Die multiple Regressionsanalyse dient der Kontrolle von Störvariablen sowie der Prognose und Erklärung des Verhaltens anhand mehrerer unabhängiger Variablen.

- (2) Wie lautet die Regressionsgleichung für Merkmalsträger?**

Die Gleichung lautet:

$$y_m = b_0 + b_1 \cdot x_{m1} + b_2 \cdot x_{m2} + \dots + b_j \cdot x_{mj} + \dots + b_k \cdot x_{mk} + e_m$$

- (3) Warum handelt es sich bei der multiplen Regressionsanalyse um ein kompensatorisches Modell?**

Die multiple Regressionsanalyse ist ein kompensatorisches Modell, da niedrige Werte auf einer unabhängigen Variablen durch hohe Werte auf anderen unabhängigen Variablen ausgeglichen werden können.

- (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?**

- (a) Das Regressionsgewicht einer Variablen in der multiplen Regression ist gleich dem Regressionsgewicht dieser Variablen in einer einfachen Regression, wenn alle Prädiktorvariablen untereinander unkorreliert sind.
- (b) Das Regressionsgewicht einer Variablen  $X_1$  in der multiplen Regression ist kleiner als ihr Regressionsgewicht in der einfachen Regression, wenn folgende Beziehung gilt:

$$r_{YX_1} > \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2}$$

- c) Das Regressionsgewicht einer Variablen  $X_1$  in der multiplen Regression ist größer als ihr Regressionsgewicht in der einfachen Regression, wenn folgende Beziehung gilt:

$$r_{YX_1} < \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r_{X_1X_2}^2}$$

- (5) Wie hängt das Partialregressionsgewicht mit der Partialkorrelation zusammen?**

Zwischen einem Partialregressionsgewicht  $b_1$  und der Partialkorrelationen  $r_{X_1Y \cdot X_2}$  besteht folgende Beziehung:

$$b_1 = r_{X_1Y \cdot X_2} \cdot \frac{\sqrt{s_Y^2 \cdot (1 - r_{X_2Y}^2)}}{\sqrt{s_{X_1}^2 \cdot (1 - r_{X_2X_1}^2)}}$$

**(6) Für welche Fragestellungen verwendet man unstandardisierte, für welche standardisierte Regressionsgewichte?**

Auf unstandardisierte Regressionsgewichte greift man beim Vergleich mehrerer Gruppen zurück; standardisierte Regressionsgewichte eignen sich zum Vergleich verschiedener Variablen.

**(7) Was bedeutet die multiple Korrelation, und wie berechnet man sie?**

Die multiple Korrelation ist die bivariate Korrelation zwischen der abhängigen Variablen und der aufgrund der unabhängigen Variablen vorhergesagten abhängigen Variablen. Sie lässt sich bspw. berechnen, indem die Quadratwurzel des Determinationskoeffizienten bestimmt wird.

**(8) Wie lautet die Quadratsummenzerlegung der multiplen Regressionsanalyse?**

Die Quadratsummenzerlegung lautet:

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

**(9) Was versteht man darunter, dass man die multiple Determination als Summe von Semipartialdeterminationen zunehmend höherer Ordnung darstellen kann?**

Veranschaulicht am Beispiel von drei unabhängigen Variablen, erhält man folgende Zerlegung:

$$R_{Y|X_1, X_2, X_3}^2 = r_{X_1 Y}^2 + r_{(X_2 \bullet X_1) Y}^2 + r_{(X_3 \bullet X_1, X_2) Y}^2$$

Dies bedeutet, dass zunächst der erste Prädiktor ( $X_1$ ) mit dem Kriterium korreliert wird. Das Quadrat der Korrelation entspricht dem Varianzanteil des Kriteriums  $Y$ , den dieser erste Prädiktor ( $X_1$ ) erklären kann. 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 ( $r_{(X_2 \bullet X_1) Y}$ ) gibt den Varianzanteil des Kriteriums an, den der zweite Prädiktor zusätzlich zum ersten Prädiktor erklärt. 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 ( $r_{(X_3 \bullet X_1, X_2) Y}$ ) gibt den Varianzanteil des Kriteriums wieder, den der dritte Prädiktor zusätzlich zu den beiden ersten Prädiktoren erklärt. Die Summe aller so berechneten Semipartialdeterminationen entspricht der multiplen Determination (dem Determinationskoeffizienten in der multiplen Regressionsanalyse).

**(10) Wie lautet das Populationsmodell der multiplen Regressionsanalyse?**

Das Populationsmodell lautet:

$$Y = E(Y|X_1, \dots, X_k) + \varepsilon = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_j \cdot X_j + \dots + \beta_k \cdot X_k + \varepsilon$$

**(11) Welche Voraussetzungen zur inferenzstatistischen Testung werden im Modell mit deterministischen Regressoren gemacht?**

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.

**(12) Welche Voraussetzungen zur inferenzstatistischen Testung werden im Modell mit stochastischen Regressoren gemacht?**

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.

**(13) Wie überprüft man inferenzstatistisch die Nullhypothese, dass**

**(a) der Determinationskoeffizient in der Population**

**(b) der quadrierte Semipartialdeterminationskoeffizient in der Population gleich 0 ist?**

(a) 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:

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

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 ( $df_1 = k$ ). Überschreitet der empirische  $F$ -Wert einen kritischen  $F$ -Wert, der zu einem a priori festgelegten  $\alpha$ -Niveau gehört, verwirft man die Nullhypothese.

(b) Zur Überprüfung dieser Hypothese gibt es einen  $F$ -Test, der wie folgt lautet:

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

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. 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). Überschreitet der empirische  $F$ -Wert einen kritischen  $F$ -Wert, der zu einem a priori festgelegten  $\alpha$ -Niveau gehört, verwirft man die Nullhypothese.

**(14) Wie lautet die Effektgröße  $\phi_1^2$  zur Bestimmung der optimalen Stichprobengröße?**

Die Berechnung des optimalen Stichprobenumfangs basiert auf der Effektgröße

$$\phi_1^2 = \frac{P_1^2}{1 - P_1^2}.$$

**(15) Welche Verfahren zur Auswahl unabhängiger Variablen kennen Sie? Erläutern Sie diese.**

Man unterscheidet theoretische und datenbasierte Auswahlstrategien. Bei der theoretischen Auswahl unterzieht man diejenigen unabhängigen Variablen einer Regressionsanalyse, von denen man aufgrund theoretischer Überlegungen eine Aufklärung der Varianz der abhängigen Variablen erwartet. In verschiedenen Forschungskontexten ist es sinnvoll, die unabhängigen Variablen zu

Gruppen zusammenzufassen und diese nacheinander blockweise in die Regressionsanalyse aufzunehmen. Die datengesteuerte Auswahl ist z. B. dann sinnvoll, wenn kein theoretisches Modell vorliegt und man sehr viele unabhängige Variablen zur Verfügung hat, von denen man die besten zur Prädiktion auswählen möchte. Dies bedeutet, dass man aus einer Menge zur Verfügung stehender Prädiktoren diejenigen auswählen möchte, 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 aufgenommen werden. Zur datengesteuerten Auswahl gibt es drei Strategien:

- (a) die Vorwärtsselektion,
- (b) die Rückwärtselimination und
- (c) die schrittweise Regression.

Bei allen drei Verfahren muss vorher ein  $\alpha$ -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.

- (a) Bei der Vorwärtsselektion gibt man dem Statistikprogramm alle unabhängigen Variablen bekannt, die für die Vorhersage des Kriteriums in Frage 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, die über die bereits in der Gleichung enthaltene Variable hinaus am meisten zusätzliche Kriteriumsvarianz erklärt, also 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 Kriteriumsvarianz 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.
- (b) Bei der Rückwärtselimination 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 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. 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.
- (c) Bei der schrittweisen Regression werden beide Strategien kombiniert. Man startet mit einer Vorwärtsselektion auf der Basis eines vorher festgelegten Signifikanzniveaus für den Einschluss ( $\alpha_E$ ). Nimmt man sukzessive neue unabhängige Variablen auf, kann es passieren, dass bei einer bestimmten Kombination unabhängiger Variablen der Vorhersagebeitrag einer bereits aufgenommen Variablen nicht länger signifikant ist. Überschreitet der  $p$ -Wert einer solchen Variablen ein vorher festgelegtes Signifikanzniveau für den Ausschluss ( $\alpha_A$ ), so wird diese Variable entfernt, bevor eine weitere neue Variable aufgenommen wird.

**(16) Was versteht man unter dem Prognosefehler?**

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, ist der Prognosefehler *PRESS* definiert als:

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

Er ist die Summe der quadrierten Abweichungen der beobachteten von den vorhergesagten Werten.

**(17) Was versteht man unter dem Kreuzvalidierungsfehler?**

Der Kreuzvalidierungsfehler *CVE* ist die Summe der quadrierten Abweichungen der beobachteten von den vorhergesagten Werten, die durch die Stichprobengröße  $n$  geteilt wird:

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

**(18) Was versteht man unter einer Suppressorvariablen?**

Eine Suppressorvariable ist eine unabhängige Variable, deren Aufnahme in das multiple Regressionsmodell dazu führt, dass der Beitrag einer anderen unabhängigen Variablen zur Erklärung der Variation der abhängigen Variablen erhöht wird.

**(19) Welche Typen von Suppressorvariablen unterscheidet man, und was bedeuten sie?**

Suppressorvariablen lassen sich in klassische, reziproke und negative Suppressorvariablen unterteilen. Unter einer klassischen Suppressorvariablen versteht man eine Variable, die mit dem Kriterium unkorreliert ist, mit einer anderen unabhängigen Variablen jedoch eine bedeutsame Korrelation aufweist. Eine reziproke Suppressorvariable ist wie eine zweite unabhängige Variable positiv mit der Kriteriumsvariablen korreliert, sie ist jedoch mit der zweiten unabhängigen Variablen negativ korreliert. Eine negative Suppressorvariable ist dadurch gekennzeichnet, dass sie mit einer zweiten unabhängigen Variablen positiv korreliert ist, die Korrelation der Suppressorvariablen mit der Kriteriumsvariablen jedoch kleiner ist als das Produkt aus der Korrelation der zweiten unabhängigen Variablen mit dem Kriterium und der Korrelation der beiden unabhängigen Variablen untereinander.

**(20) Was versteht man unter der Nützlichkeit einer Variablen?**

Die Nützlichkeit einer unabhängigen Variablen gibt an, wie viel Varianz der abhängigen Variablen diese unabhängige Variable zusätzlich zu allen anderen unabhängigen Variablen erklärt. Sie entspricht der Semipartialdetermination der höchstmöglichen Ordnung.

**(21) Was versteht man unter Zentrierung von Variablen, und weswegen nimmt man sie vor?**

Die Zentrierung bedeutet, dass man von jedem Messwert den Mittelwert der Variablen abzieht. Dies führt dazu, dass jede zentrierte Variable einen Mittelwert von 0 hat. Man nimmt die Zentrierung von Variablen u. a. in der moderierten Regressionsanalyse vor, um die Multikollinearität zu verringern und die Interpretation zu erleichtern.

**(22) Beschreiben Sie das Vorgehen zur Überprüfung von Interaktionseffekten in der multiplen Regressionsanalyse (moderierte Regression).**

Um Interaktionseffekte zu überprüfen, bildet man zunächst Interaktionsterme, indem man die unabhängigen Variablen, zwischen denen eine Interaktion vermutet wird, multipliziert und das Produkt als weitere unabhängige Variable in die Regressionsgleichung aufnimmt. Vor der Multiplikation werden die unabhängigen Variablen zentriert, um die Multikollinearität zu verringern und die Interpretation zu erleichtern. Zur statistischen Überprüfung des Interaktionseffekts überprüft man die Nullhypothese, dass das Regressionsgewicht der Produktvariablen in der Population gleich 0 ist. Hierzu kann man auf eine  $t$ -verteilte Prüfgröße zurückgreifen, bei der das geschätzte Regressionsgewicht durch seinen Standardfehler geteilt wird.

**(23) Was sind bedingte Regressionsgewichte?**

Das bedingte Regressionsgewicht ist das Regressionsgewicht einer unabhängigen Variablen gegeben die Ausprägungen der anderen unabhängigen Variablen.

**(24) Wie kann man mit der Regressionsanalyse nicht-lineare Zusammenhänge untersuchen?**

Nicht-lineare Zusammenhänge lassen sich durch die Hinzunahme von Polynomen höherer Ordnung berücksichtigen.

**(25) Was versteht man unter Dummy-Codierung, und nach welchem Prinzip werden die Codiervariablen gebildet?**

Eine Dummy-Codierung wird benötigt, wenn kategoriale unabhängige Variablen in eine multiple Regressionsanalyse aufgenommen werden sollen. Die Dummy-Codierung erfolgt in mehreren Schritten. Zunächst wird eine der Kategorien der unabhängigen Variablen als Referenzkategorie ausgewählt. Dieser Referenzkategorie wird auf allen Codiervariablen der Wert 0 zugewiesen. Allen anderen Kategorien der unabhängigen Variablen werden Werte auf den Codiervariablen derart zugewiesen, dass (a) jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0, und (b) jede Codiervariable nur für eine einzige Kategorie den Wert 1 aufweist, für alle anderen den Wert 0.

**(26) Worin unterscheiden sich die ungewichtete und die gewichtete Effektcodierung?**

Bei der ungewichteten Effektcodierung geht man so vor, dass zunächst eine Kategorie der unabhängigen Variablen als Referenzkategorie ausgewählt wird. Der Referenzkategorie wird auf allen Codiervariablen der Wert  $-1$  zugewiesen. Allen anderen Kategorien der unabhängigen Variablen wird ein Wert auf den Codiervariablen derart zugewiesen, dass jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0 und 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. Dies hat zur Folge, dass der Achsenabschnitt dem ungewichteten Mittelwert der Kategorienmittelwerte entspricht, wohingegen ein Regressionsgewicht die Differenz zwischen dem Mittelwert einer Kategorie und dem ungewichteten Mittelwert der Kategorienmittelwerte repräsentiert.

Bei einer gewichteten Effektcodierung bildet man die Werte der Codiervariablen derart, dass die Regressionsgewichte nicht Abweichungen vom ungewichteten Mittelwert der Kategorien, sondern Abweichungen vom gewichteten Mittelwert der Kategorien (also dem Gesamtmittelwert) darstellen.

len. Der Achsenabschnitt entspricht dem gewichteten Mittelwert. Der gewichtete Mittelwert wird gebildet, indem die Mittelwerte in den Kategorien mit ihrer relativen Häufigkeit gewichtet und aufsummiert werden. Die Wertezuweisung auf den Codiervariablen erfolgt so, der Referenzkategorie auf der Codiervariablen  $X_j$  der Wert  $-(n_{X_j}/n_R)$  zugewiesen wird, wobei  $n_{X_j}$  die Stichprobengröße derjenigen Kategorie ist, der auf der Codiervariablen  $X_j$  der Wert 1 zugewiesen wird, und  $n_R$  die Stichprobengröße der Referenzkategorie bezeichnet. Allen anderen Kategorien der unabhängigen Variablen wird ein Wert auf den Codiervariablen derart zugewiesen, dass jede Kategorie nur auf einer einzigen Codiervariablen einen Wert von 1 aufweist, auf allen anderen Codiervariablen den Wert 0, und 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.

**(27) Was bedeuten die Regressionskoeffizienten bei der Dummy-, was bei der Effektcodierung?**

Bei der Dummy-Codierung entspricht der Achsenabschnitt  $b_0$  dem Mittelwert der Variablen in der Referenzkategorie. Das Regressionsgewicht  $b_j$  ist die Differenz der Mittelwerte der Kategorie  $j$  und der Referenzkategorie. Bei der ungewichteten Effektcodierung entspricht der Achsenabschnitt  $b_0$  dem ungewichteten Mittelwert der Kategorienmittelwerte. Das Regressionsgewicht  $b_j$  ist die Differenz zwischen dem Mittelwert der Kategorie  $j$  und dem ungewichteten Gesamtmittelwert. Bei der gewichteten Effektcodierung entspricht der Achsenabschnitt  $b_0$  dem gewichteten Mittelwert der Kategorienmittelwerte. Das Regressionsgewicht  $b_j$  ist die Differenz zwischen dem Mittelwert der Kategorie  $j$  und dem gewichteten Mittelwert der Kategorienmittelwerte (also dem Gesamtmittelwert).

**(28) Wie überprüft man Interaktionen zwischen kategorialen unabhängigen Variablen mit der multiplen Regressionsanalyse?**

Um Interaktionen zwischen kategorialen 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 Kodierung der Bedingungen der ersten unabhängigen Variablen benötigt werden, mit allen Codiervariablen, die zur Kodierung der Bedingungen der zweiten unabhängigen Variablen benötigt werden, multipliziert. Zur statistischen Überprüfung greift man dann auf einen  $F$ -Test zurück, bei dem der Determinationskoeffizient des uneingeschränkten Modells, das alle unabhängigen Variablen enthält, mit dem Determinationskoeffizienten des eingeschränkten Modells, das alle unabhängigen Variablen bis auf die Produktvariablen enthält, verglichen wird. Unterscheiden sich beide Determinationskoeffizienten signifikant voneinander, verwirft man die Nullhypothese, dass keine Interaktion vorliegt.

**(29) Was ist eine Kovarianzanalyse?**

Bei der Kovarianzanalyse werden metrische und kategoriale Variablen als unabhängige Variable berücksichtigt. Es wird angenommen, dass sie additiv, nicht aber multiplikativ zusammenwirken.

**(30) Was versteht man unter adjustierten Mittelwerten, und warum bestimmt man sie?**

Adjustierte Mittelwerte sind die mittels der Kovarianzanalyse vorhergesagten Gruppenmittelwerte an der Stelle der Gesamtmittelwerte der unabhängigen Variablen. Adjustiert bedeutet, dass Personen gleicher Ausprägung auf den Kovariaten betrachtet und dadurch Unterschiede zwischen Per-



sonen konstant halten werden, die auf die Kovariaten zurückgeführt werden können. Die adjustierten Mittelwerte geben also die mittleren Unterschiede zwischen den Gruppen wieder, die nicht auf Unterschiede in den Kovariaten zurückgeführt werden können.

**(31) Welche Probleme stellen sich, wenn man die Kovarianzanalyse zur Auswertung quasi-experimenteller Untersuchungspläne in der Evaluationsforschung einsetzt?**

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. Wird der Messfehler nicht berücksichtigt, können sich bedeutsame Scheineffekte zeigen, d. h., Unterschiede in der abhängigen Variablen, die zwischen den Ausprägungen der kategorialen unabhängigen Variablen bestehen, können fälschlicherweise kausal interpretiert werden. Auch können ausgelassene Drittvariablen, die mit der kategorialen unabhängigen Variablen konfundiert sind, dazu führen, dass sich Scheineffekte der kategorialen Variablen zeigen, die wiederum fälschlicherweise als kausale Effekte interpretiert werden könnten.

**(32) Was ist Lords Paradox?**

Lords Paradox beschreibt das Phänomen, dass die Analyse von mittleren Veränderungen und die Kovarianzanalyse zu einander widersprechenden Ergebnissen in Bezug auf Gruppeneffekte führen können, wenn Veränderungen für natürlich vorgefundene, d. h. nicht randomisiert gebildete Gruppen in einem Vortest-Nachtest-Design untersucht werden.

**(33) Was versteht man unter einer Aptitude-Treatment-Interaction-Analyse?**

Bei der Aptitude-Treatment-Interaction-Analyse untersucht man die regressive Abhängigkeit einer abhängigen Variablen von kategorialen und metrischen unabhängigen Variablen. Dabei werden Interaktionen zwischen den metrischen und den kategorialen unabhängigen Variablen zugelassen.

**(34) Was versteht man unter Under- und was unter Overfitting, und welche Konsequenzen haben diese?**

Das Auslassen relevanter unabhängiger Variablen in der multiplen Regressionsanalyse nennt man Underfitting, die Aufnahme von irrelevanten unabhängigen Variablen Overfitting.

Das Underfitting hat zur Folge, dass die Regressionskoeffizienten verzerrt geschätzt werden. Darüber hinaus geht das Auslassen relevanter unabhängiger Variablen mit einem Verlust an Teststärke einher. Die Hinzunahme irrelevanter unabhängiger Variablen (Overfitting) kann zu einer verzerrten Schätzung der Regressionsgewichte der anderen Variablen führen, wodurch Prognose- und Kreuzvalidierungsfehler begünstigt werden.

**(35) Was ist ein LOWESS-Anpassungsverfahren?**

Durch ein LOWESS-Anpassungsverfahren kann eine Linie in ein bivariates 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 spezielles Glättungsverfahren.



**(36) Wie kann man Ausreißerwerte auf der abhängigen und der unabhängigen Variablen identifizieren?**

Bei der Identifikation von Ausreißerwerten auf der abhängigen Variablen greift man auf die Residuen zurück. Insbesondere sollte das studentisierte ausgeschlossene Residuum zur Bestimmung von Ausreißerwerten herangezogen werden. Beim studentisierten Residuum wird das Residuum durch die geschätzte Standardabweichung des Residuums an einer Stelle der unabhängigen Variablen dividiert. Beim studentisierten ausgeschlossenen Residuum wird die betrachtete Person 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 einfließt. 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.

Um Ausreißerwerte auf einer unabhängigen Variablen aufzudecken, kann man auf die Mahalanobis-Distanz und die Hebelwerte zurückgreifen. Die Mahalanobisdistanz ist wie folgt definiert:

$$d_m = \sqrt{\frac{(x_m - \bar{x})^2}{\hat{\sigma}_x^2}}. \text{ In Bezug auf die Hebelwerte greift man auf zentrierte Hebelwerte zurück, die}$$

der quadrierten Mahalanobis-Distanz geteilt durch  $(n - 1)$  entsprechen. Die zentrierten Hebelwerte können Werte zwischen 0 und  $1 - 1/n$  annehmen. Zur Bewertung dieser Hebelwerte werden in der Literatur verschiedene Schwellenwerte diskutiert, ab denen Hebelwerte als auffällig gelten. Z. B. werden als untere Schwellenwerte  $2 \cdot k/n$  bei großen und  $3 \cdot k/n$  bei kleinen Stichproben genannt. Eine andere Empfehlung besteht darin, sich die Verteilung der Hebelwerte anzuschauen und nur diejenigen Werte genauer zu inspizieren, die sich stark von den anderen unterscheiden.

**(37) Was sind einflussreiche Datenpunkte, und wie kann man sie identifizieren?**

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. 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. 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 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. Um die Veränderungen in Bezug auf verschiedene Regressionskoeffizienten vergleichbar zu machen, werden DfBETAS-Werte 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. DfBETAS-Werte, die in kleinen bzw. mittelgroßen Stichproben vom Betrag her größer als 2 sind, zei-

gen auffällige Werte an, für große Stichproben wird empfohlen, DfBETAS-Werte kritisch zu betrachten, deren Betrag größer als  $2/\sqrt{n}$  ist.

Um die Frage zu untersuchen, wie sich die erwarteten  $\hat{y}$ -Werte ändern, wenn man eine Person der Stichprobe entnimmt, 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 in 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.

Neben den DfFITS-Werten kann auch Cooks Distanz herangezogen werden, um einflussreiche Datenpunkte in Bezug auf die vorhergesagten Werte zu identifizieren. Für Cooks Distanz gibt es kritische Schwellenwerte, die auf einer F-Verteilung aufbauen.

**(38) Was versteht man unter Multikollinearität, und wie kann man sie aufdecken?**

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. Den Toleranzfaktor erhält man, indem man die quadrierte multiple Korrelation einer unabhängigen Variablen mit allen anderen unabhängigen Variablen von 1 abzieht. 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. Der Varianzinflations-Faktor ist der Kehrwert der Toleranz. Ein Wert des Varianzinflations-Faktors, der größer als 10 ist, wird in der Literatur häufig als auffallend bewertet.

**(39) Was sind die Konsequenzen von Heteroskedastizität?**

Heteroskedastizität führt zu verzerrten Standardfehlern.

**(40) Was ist ein Residuenplot, und wofür setzt man ihn ein?**

In einem Residuenplot werden Residuen, üblicherweise studentisierte Residuen, auf der Y-Achse gegen die aufgrund der Regression vorhergesagten  $\hat{y}$ -Werte auf der X-Achse abgebildet. Mit Residuenplots können Verletzungen der Annahme der Regressionsanalyse wie bspw. Verletzungen der Homoskedastizität und Fehlspezifikationen aufgedeckt werden.

**(41) Was versteht man unter der Unabhängigkeit von Residuen?**

Unabhängigkeit der Residuen bedeutet, dass die Residuen voneinander stochastisch unabhängig sind. Diese Annahme ist in zwei typischen Anwendungsfällen in der Psychologie verletzt: (1) wenn der Stichprobenziehung Klumpenstichproben oder mehrstufige Auswahlverfahren zugrunde liegen, (2) bei serialer Abhängigkeit, die typischerweise in Einzelfalluntersuchungen auftritt.

**(42) Wie kann man die Normalverteilungshypothese untersuchen?**

Die bedingte Normalverteilungsannahme der Residuen kann mit einem Histogramm und einem P-P-Plot der studentisierten Residuen graphisch überprüft werden. Es wird empfohlen, ein Histogramm der studentisierten Residuen zu erstellen, das dann in Bezug auf Abweichungen von der Normalverteilung inspiziert werden kann. Beim Probability-Probability-Plot (P-P-Plot) werden zwei kumulierte Wahrscheinlichkeiten gegeneinander abgetragen. Auch für diesen Plot sollte man auf die studentisierten Residuen zurückgreifen. Auf der Abszisse werden 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 werden die kumulierten Wahrscheinlichkeiten für einen spezifischen Abszissenwert abgetragen, die man bei Gültigkeit eines bestimmten Verteilungsmodells erwarten würde. Sind die Residuen normalverteilt, liegen alle Punkte auf einer Geraden. Gravierende Abweichungen von der Geraden weisen auf eine Verletzung der Normalverteilungsannahme hin.

**(43) Welche Konsequenzen haben die Verletzungen von Annahmen, auf denen die multiple Regressionsanalyse basiert?**

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. Des Weiteren wird Homoskedastizität angenommen. Eine Verletzung der Homoskedastizität führt zu verzerrten Standardfehlern. Verletzungen der Normalverteilungsannahme führen bei kleinen Stichproben zu verzerrten Schätzungen der Standardfehler. Verletzungen der Unabhängigkeit der Residuen führen zu verzerrten Schätzungen der Standardfehler. Hohe Multikollinearität führt zu hohen Standardfehlern der Regressionsgewichte.