24 Lineare Regressionsanalyse

Christof Wolf und Henning Best

[a] GESIS – Leibniz-Institut für Sozialwissenschaften und Universität Mannheim
[b] Universität Mannheim

Zusammenfassung. Die Regressionsanalyse ist ein Verfahren zur Schätzung des Einflusses einer oder mehrerer Merkmale auf eine abhängige Variable. Der große Vorteil der Regressionsanalyse ist, dass sie den Einfluss eines einzelnen Merkmals auf eine abhängige Variable unter Konstanthaltung der anderen Einflussgrößen schätzt. Bei der linearen Regression werden nur lineare bzw. linearisierbare Einflussbeziehungen auf metrisch abhängige Variablen erfasst. Auf der Grundlage eines Beispiels führt dieser Beitrag zunächst das Grundprinzip der linearen Regression ein. Im zweiten Abschnitt werden die mathematisch-statistischen Grundlagen des Verfahrens genauer beschrieben. Ausführlich dargestellt werden das Modell der linearen Regression, die Methode der kleinsten Quadrate, die Bestimmung der Modellgüte, die statistische Absicherung der Modellergebnisse und ihre Interpretation. Es folgt die Diskussion eines ausführlichen Beispiels, anhand dessen die wichtigsten Analysestrategien und Prinzipien der Regressionsanalyse erläutert werden. Abschließend geben wir Hinweise zu den typischen Fehlern, die bei der Anwendung des Verfahrens gemacht werden können, und empfehlen weiterführende Literatur.

1 Einführung

Regressionsanalytische Verfahren gehören heute in den Sozialwissenschaften zu den am häufigsten verwendeten Auswertungsverfahren. Allen regressionsanalytischen Verfahren ist gemeinsam, dass mit ihnen überprüft werden kann, inwieweit ein interessierendes Merkmal auf andere Merkmale „zurückgeführt“ werden kann. Hier wird denn auch der lateinische Ursprung der Bezeichnung „Regressions“-analyse deutlich, die sich von *regredi* (zurückgehen) oder *regressio* (die Rückkehr) ableitet. Typische Fragestellungen, bei der die Regressionsanalyse eingesetzt werden könnte, lauten: Wie stark ist der Einfluss der Berufserfahrung auf das Einkommen? Welche Faktoren beeinflussen die Lebenszufriedenheit? Hat eine Zunahme des Umweltwissens eine Veränderung des Umweltverhaltens zur Folge?

Das Merkmal, welches jeweils erklärt werden soll, wird auch als abhängige Variable bezeichnet; in den genannten Beispielen wären dies das Einkommen, die Lebenszufriedenheit und das Umweltverhalten. Die erklärenden Merkmale werden dementsprechend als unabhängige Variablen oder als Prädiktoren bezeichnet. Dabei ist die Einteilung in unabhängige und abhängige Variablen immer im Zusammenhang mit einer konkreten Fragestellung zu sehen. Bei anderen Fragestellungen kann die Zuordnung anders erfolgen. So ist das Umweltwissen im oben genannten Beispiel eine unabhängige Variable,

S. 607–638 in: Christof Wolf & Henning Best, Hg. (2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften

deren Einfluss auf das Umweltverhalten untersucht wird. In einem weiteren Schritt könnte untersucht werden, von welchen Faktoren das Umweltwissen seinerseits abhängt. Die abhängige Variable wäre dann das Umweltwissen.

Je nach Skalenniveau der abhängigen Variablen kommen unterschiedliche Varianten der Regressionsanalyse in Frage. Für binäre abhängige Variablen kann die logistische Regressionsanalyse verwendet werden (vgl. Kapitel 31 in diesem Handbuch), für nominalskalierte Merkmale mit mehr als zwei Ausprägungen und für ordinalskalierte Merkmale stehen verallgemeinerte Varianten der logistischen Regressionsanalyse zur Verfügung (vgl. Kapitel 32 in diesem Handbuch). Für Zähldaten wird dagegen oft auf die Poissonregression zurückgegriffen (vgl. Kapitel 33 in diesem Handbuch). In diesem Kapitel werden die Grundlagen der linearen Regressionsanalyse dargestellt, die zur Voraussetzung hat, dass die abhängige Variable metrisch skaliert ist. Zunächst sollen die Grundzüge der linearen Regressionsanalyse am Beispiel eines Modells zur Untersuchung des Einkommens abhängig Beschäftigter dargestellt werden. Eine ausführliche Beschreibung des Erklärungsmodells und der verwendeten Merkmale erfolgt in Abschnitt 3 dieses Kapitels.

Die erste zu untersuchende Hypothese sei, dass das Einkommen mit zunehmender Berufserfahrung steigt. Das Einkommen ist die abhängige Variable, das Merkmal, das mit Hilfe des statistischen Modells untersucht werden soll. Die Berufserfahrung ist die unabhängige Variable, also das Merkmal, dessen Einfluss auf das Einkommen hier geprüft wird. Mathematisch lässt sich dies in der Gleichung

$$\text{Einkommen} = f(\text{Berufserfahrung})$$

audrücken. Diese Schreibweise bringt zum Ausdruck, dass wir davon ausgehen, das Einkommen sei eine Funktion der Berufserfahrung. Dabei bleibt zunächst offen, welcher Art diese Funktion ist. Wird vermutet, dass das Einkommen mit jedem Berufsjahr um einen konstanten Betrag ansteigt, kann dies mit der Funktion

$$\text{Einkommen} = \beta_0 + \beta_1\,\text{Berufserfahrung} + \text{Fehlerterm}$$

zum Ausdruck beschrieben werden. Neben den Variablen Einkommen und Berufserfahrung enthält diese Gleichung zwei sog. Regressionskoeffizienten oder Parameter, β_0 und β_1. Außerdem taucht noch eine als „Fehlerterm“ bezeichnete Größe auf. Diese bringt die Vermutung zum Ausdruck, dass es sich bei dem Zusammenhang zwischen Berufserfahrung und Einkommen nicht um eine deterministische (perfekte) funktionale Beziehung handelt. Vielmehr werden auch andere Faktoren das Einkommen beeinflussen, einige von ihnen systematisch, andere werden zu einer zufälligen Schwankung des Einkommens beitragen. In eine mathematische Notation überführt, lautet unser Modell

$$y = \beta_0 + \beta_1 x + \varepsilon\,. \tag{1}$$

Gleichung (1) gibt das Grundmodell der bivariaten Regressionsanalyse wieder. Wenn wir den Fehlerterm ε auf beiden Seiten dieser Gleichung subtrahieren, ergibt sich

$$y - \varepsilon = \hat{y} = \beta_0 + \beta_1 x\,, \tag{2}$$

wobei mit $\hat{y}$ (y-Dach) die auf Basis der x-Werte geschätzten y-Werte bezeichnet werden. Wie Gleichung (2) zeigt, stehen x und die auf Basis von x geschätzten Werte $\hat{y}$ in einer linearen Beziehung zueinander, das heißt, alle Wertepaare (x,$\hat{y}$), die Gleichung (2) erfüllen, liegen auf einer Geraden.

Betrachten wir nun die Regressionskoeffizienten β_0 und β_1 genauer. In dem Modell, das durch Gleichung (1) spezifiziert wird, wird davon ausgegangen, dass die Berufserfahrung sich auf das Einkommen auswirkt, und zwar derart, dass sich das Einkommen um β_1 Einheiten verändert, wenn die Berufserfahrung um eine Einheit, z. B. ein Jahr, steigt. Daher wird der Koeffizient β_1 auch als Steigungskoeffizient bezeichnet. Der Regressionskoeffizient β_0 gibt den so genannten y-Achsenabschnitt an, also den Wert, bei dem die Regressionsgerade die y-Achse schneidet. In dem von uns gewählten Beispiel entspricht β_0 dem geschätzten Einkommen von Personen ohne Berufserfahrung, genauer: von Personen, bei denen das Merkmal Berufserfahrung den Wert null annimmt.

Nachdem für eine gegebene Fragestellung ein entsprechendes Regressionsmodell spezifiziert wurde, besteht der nächste Schritt darin, dieses Modell anhand empirischer Daten zu schätzen. Wir haben eine entsprechende Analyse auf Basis des ALLBUS 2006 durchgeführt. Als abhängige Variable verwenden wir das persönliche monatliche Nettoeinkommen in Euro, als unabhängige Variable die Berufserfahrung in Jahren.[1] Schätzt man mit Hilfe dieser Merkmale das in Gleichung (1) wiedergegebene Modell, erhält man für die Gruppe der abhängig Beschäftigten in Vollzeit folgendes Resultat:

$$\widehat{\text{Nettoeinkommen}} = 1371 + 18{,}4 \cdot \text{Berufserfahrung}\,.$$

Der Achsenabschnitt (β_0) beträgt also 1371 €, der Steigungskoeffizient (β_1) für die Berufserfahrung 18,4 €. Demnach verdienen Berufsanfänger, d. h. Beschäftigte ohne Berufserfahrung, durchschnittlich 1371 €. Obwohl sich unsere Vermutung, dass das Einkommen mit zunehmender Berufserfahrung ansteigt, bestätigt (der Steigungskoeffizient ist positiv), erscheint die mit jedem Berufsjahr durchschnittlich erfolgende Steigerung um 18,4 € gemessen am durchschnittlichen „Anfangsgehalt“, als gering. Im Vergleich zu einem Berufsanfänger verdient ein Beschäftigter mit 45 Berufsjahren im Durchschnitt „nur“ 828 € mehr.

Eine konkretere Vorstellung über die Art des untersuchten Zusammenhangs vermittelt Abbildung 1. Die Abweichungen zwischen den beobachteten Werten (die Punkte im Streudiagramm) und den vorhergesagten Werten auf der Regressionsgeraden sind relativ groß. Das Ausmaß dieser Abweichungen lässt sich numerisch mit dem Koeffizienten R^2 bestimmen. Diese Maßzahl gibt an, welcher Anteil der beobachteten Varianz – also der Einkommensunterschiede – durch das Regressionsmodell reproduziert werden kann (eine genauere Erläuterung dieser Maßzahl findet sich im nächsten Abschnitt). In unserem Anwendungsbeispiel kann die Berufserfahrung lediglich 5,8 Prozent der Varianz in den Einkommen abhängig Beschäftigter statistisch erklären ($R^2 = 0{,}058$). Der Zusammenhang zwischen Berufserfahrung und Einkommen ist demnach nur verhältnismäßig schwach ausgeprägt. Oder anders ausgedrückt: für andere Faktoren, die bisher nicht berücksichtigt wurden – die Ausbildung, das Geschlecht, die berufliche

[1] Genauere Angaben zur Operationalisierung geben wir im Abschnitt 3 dieses Beitrags.

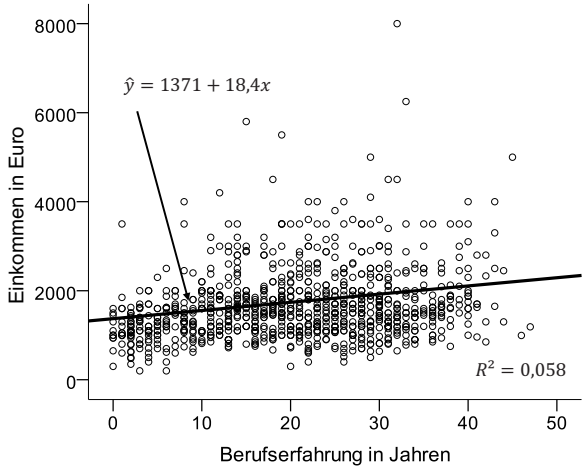
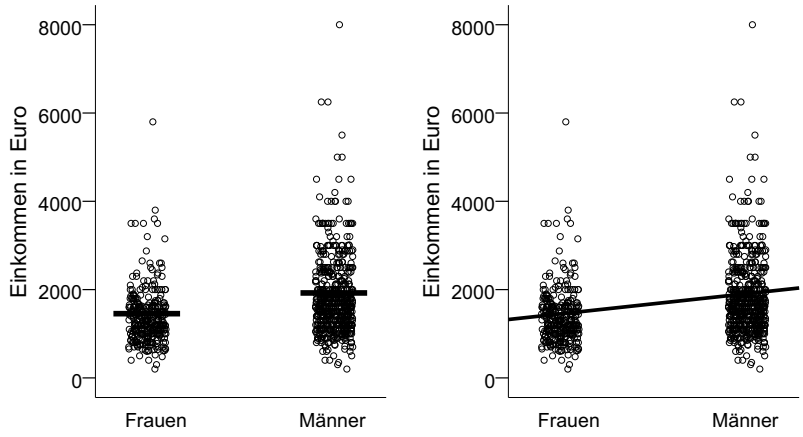


Abb. 1: Streudiagramm mit Regressionsgerade

Position usw. –, bleiben 94 % der Varianz zu erklären. Allerdings ist die Bewertung der Modellgüte, wie alle Bewertungen, normativ und nur im Hinblick auf einen Bezugspunkt sinnvoll möglich. Ein solcher Bezugspunkt könnte beispielsweise ein anderes Modell für dieselbe abhängige Variable oder das Ergebnis für dasselbe Modell aus früheren Erhebungen sein. Die Verwendung solch empirischer Bezugspunkte erscheint uns angemessener als vorgegebene Daumenregeln, nach denen R^2-Werte bis zu einer gewissen Größe als schwach, dann als mittel und schließlich als stark gelten können. In jedem Fall kann die Bewertung der Modellgüte nur in Relation zur untersuchten Fragestellung beantwortet werden.

Wie erläutert, gibt der Anteil erklärter Varianz Auskunft über die Güte des untersuchten Modells. Eine davon unabhängige Frage betrifft die Stärke des untersuchten Einflusses. Diese wird durch den Steigungskoeffizienten ausgedrückt. In unserem Beispiel beträgt die Steigung und damit die Effektstärke 18,4 € je Berufsjahr, d. h. 184 € in 10 Berufsjahren. Um zu beurteilen, ob es sich dabei um einen großen oder kleinen Effekt handelt, hilft noch einmal ein Blick auf Abbildung 1. Relativ zur Skala, auf der das Einkommen beobachtet wird, fällt die jährliche Steigerung von 18,4 € klein aus, die Steigung der Regressionsgeraden ist eher gering. Wie bei allen statistischen Verfahren kann die inhaltliche Interpretation der Regressionsanalyse jedoch nicht allein auf statistischen Kriterien beruhen. Diese muss vielmehr vor dem Hintergrund theoretischer Annahmen und dem Wissen um relevante Randbedingungen geschehen.

Soweit haben wir als unabhängiges Merkmal eine metrische Variable verwendet. Die Regressionsanalyse bietet jedoch auch die Möglichkeit, kategoriale Prädiktoren zu analysieren. Um dies deutlich zu machen, wollen wir untersuchen, ob sich das Einkommen von Männern und Frauen unterscheidet. Auf Basis des ALLBUS 2006 ergibt sich für alle Vollzeit abhängig Beschäftigten ein durchschnittliches monatliches Nettoeinkommen von 1755 €. Allerdings unterscheidet sich das Einkommen von Män-



(a) Mittelwerte (b) Regressionsgerade

Abb. 2: Einkommensverteilung nach Geschlecht

nern und Frauen stark: Frauen kommen monatlich auf 1445 €, während Männer 1913 € verdienen, also 468 € mehr. Grafisch ist die Verteilung der Einkommen für Männer und Frauen in Abbildung 2 a dargestellt.[2] Hier zeigt sich deutlich, dass Einkommen oberhalb 2000 € bei Frauen viel seltener sind als bei Männern.

Wie können nun qualitative Merkmale, wie das Geschlecht, in die Regressionsanalyse aufgenommen werden? Dies geschieht unter Verwendung so genannter Dummy-Variablen, also Stellvertreter. Im Fall eines binären qualitativen Merkmals, wie dem Geschlecht, verkodet man in der Dummy-Variablen eine der beiden Kategorien mit null, die andere mit eins. Die mit null kodierte Kategorie wird auch als Referenzkategorie bezeichnet. Für die Regressionsgleichung ergibt sich

$$\hat{y} = \beta_0 + \beta_1 D_G \,, \tag{3}$$

wobei D_G die Dummy-Variable für Geschlecht ist und hier für Frauen mit null und für Männer mit eins kodiert wurde. Um zu verstehen, was diese Gleichung bedeutet, ist es hilfreich, sie in zwei separate Gleichungen zu schreiben, je eine für die beiden Zustände der Dummy-Variable. Für Frauen ($D_G = 0$) reduziert sich die Regressionsgleichung zu

$$\hat{y}_F = \beta_0 \,,$$

während für Männer ($D_G = 1$)

$$\hat{y}_M = \beta_0 + \beta_1$$

gilt. Der Achsenabschnitt β_0 entspricht dem erwarteten durchschnittlichen Einkommen der Frauen und der Koeffizient β_1 entspricht der Differenz zwischen dem Erwartungswert des Einkommens für Männer und Frauen. Die Vorhersagewerte $\hat{y}$ für jede

[2] Die horizontale Streuung der Punkte innerhalb der beiden Gruppen hat inhaltlich keine Bedeutung und dient lediglich dazu, die Form der Verteilung besser deutlich machen zu können.

Kategorie der Dummy-Variablen ergeben sich in der bivariaten Regression aus dem Mittelwert der jeweiligen Kategorie. Dies wird auch in Abbildung 2 b deutlich: Die Regressionsgerade verbindet die Mittelwerte der beiden untersuchten Gruppen.

Bisher haben wir nur Regressionsanalysen mit jeweils einer unabhängigen Variablen betrachtet. Man spricht in diesem Zusammenhang auch von bivariater Regressionsanalyse. Der große Nutzen aller Regressionsverfahren besteht nun aber darin, dass mehr als eine unabhängige Variable gleichzeitig in ein entsprechendes Modell aufgenommen werden kann. Die Effekte aller unabhängigen Variablen werden dann gleichzeitig geschätzt, *jeweils unter Kontrolle aller anderen unabhängigen Variablen.* Im Gegensatz zu bivariater spricht man dann von multipler Regressionsanalyse. Der nächste Schritt der Analyse könnte z. B. darin bestehen, die beiden oben diskutierten bivariaten Modelle zu einem Modell der multiplen Regression zusammenzufassen. Dieses Modell enthielte die beiden unabhängigen Variablen Berufserfahrung und Geschlecht. Formal ändert sich an der Regressionsgleichung nur, dass eine weitere Variable rechts vom Gleichheitszeichen steht, also:

$$\text{Einkommen} = \beta_0 + \beta_1\ \text{Berufserfahrung} + \beta_2\,\text{Geschlecht} + \text{Fehlerterm}$$

bzw.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_G + \varepsilon\,. \tag{4}$$

Dieses Modell ermittelt den Einfluss der Berufserfahrung unter Berücksichtigung des Geschlechts. Einen solchen „korrigierten“ Effekt bezeichnet man auch als partiellen Effekt. Gleichzeitig schätzt das Modell, wie der geschlechtsspezifische Einkommensunterschied wäre, *wenn Männer und Frauen dieselbe Berufserfahrung hätten.*

In diesem multiplen Regressionsmodell entspricht der Achsenabschnitt dem bedingten Erwartungswert der Referenzkategorie – hier also dem zu erwartenden Durchschnittseinkommen von Frauen, wenn diese dieselbe Berufserfahrung hätten wie Männer – und der Koeffizient der mit eins kodierten Kategorie(n) entspricht der Differenz der bedingten Erwartungswerte zwischen dieser Kategorie und der Referenzkategorie – in unserem Beispiel also der erwarteten Einkommensdifferenz zwischen Männern und Frauen bei gleicher Berufserfahrung. Auf der Basis dieses Modells könnte zum Beispiel ermittelt werden, ob Männer nur deshalb so viel mehr verdienen als Frauen, weil sie über mehr Berufserfahrung verfügen (siehe dazu Abschnitt 3). Bevor wir mit diesem Beispiel weiter fortfahren, soll im nächsten Abschnitt zunächst systematisch in die mathematisch-statistischen Grundlagen der multiplen Regressionsanalyse eingeführt werden.

2 Mathematisch-statistische Grundlagen

2.1 Das allgemeine Modell

Wie bereits zu Anfang dieses Aufsatzes erwähnt, ist die Regressionsanalyse ein statistisches Verfahren, mit welchem der Einfluss eines oder mehrerer Merkmale auf ein anderes Merkmal untersucht werden kann. Mathematisch lässt sich dies als

$$y = f(x_1,x_2,\ldots,x_k) + \varepsilon \tag{5}$$

formulieren. Die Formel macht deutlich, dass wir nicht von einer deterministischen Beziehung zwischen x_j und y ausgehen. Stattdessen wird eine statistische Beziehung unterstellt, bei der die unabhängigen Variablen die abhängige Variable nur mehr oder weniger gut „voraussagen“ oder „erklären“ können und in jedem Fall ein „Rest“ bleibt, der hier mit dem Symbol ε bezeichnet wird. Diese Größe wird auch Fehlerterm, Residuum oder Störgröße genannt.

Die unterschiedlichen Ansätze der Regressionsanalyse, wie sie auch im vorliegenden Band dargestellt werden (siehe die folgenden Kapitel in diesem Handbuch), unterscheiden sich danach, welches Skalenniveau die abhängige Variable y hat. Je nachdem, welches Skalenniveau für die abhängige Variable angenommen wird, wird sich die Wahl der Funktion $f(\cdot)$, also der unterstellte funktionale Zusammenhang zwischen unabhängigen und abhängigen Variablen unterscheiden. In diesem Kapitel stellen wir die lineare Regressionsanalyse vor. Dies bedeutet, dass die Funktion, die die abhängige Variable mit den unabhängigen Variablen verknüpft, linear, genauer: in den Parametern linear sein muss. Die entsprechende Gleichung lautet folglich

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon = \sum_{j=0}^{k} \beta_j x_{ij} + \varepsilon$$

mit $x_{.0} = 1$ oder in Matrixnotation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

bzw.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

Wie Gleichung (6) zeigt, ist y über eine lineare Funktion mit den x_j verbunden. Die Koeffizienten β_j werden als Regressionskoeffizienten bezeichnet; β_0 auch als Achsenabschnitt (englisch: *intercept*) und die übrigen β_j als Steigung (englisch: *slope*). Bei $\mathbf{X}\boldsymbol{\beta}$ handelt es sich um die auf Basis des Modells vorhergesagten y-Werte $\hat{\mathbf{y}}$.

2.2 Die Identifikation der Regressionskoeffizienten

Das zentrale Problem jeder Regressionsanalyse besteht darin, Schätzer für die Regressionskoeffizienten β_j, die Parameter des Regressionsmodells, so zu bestimmen, dass die vom Modell geschätzten Werte $\hat{y}$ den beobachteten Werten y möglichst gut entsprechen. Zur Lösung dieser Aufgabe existieren verschiedene Verfahren. Im Folgenden stellen wir die Methoden der kleinsten Quadrate vor, bei der es sich um das Standardverfahren für die lineare Regression handelt. Alternativ könnte auch das in Kapitel 10 dieses Handbuchs beschriebene Schätzverfahren der Maximum-Likelihood-Methode

verwendet werden. Betrachten wir noch einmal Gleichung (6), die Basisgleichung der multiplen linearen Regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$$

Wie bereits erwähnt, sollen die β_j so bestimmt werden, dass die vom Modell geschätzten $\hat{y}$-Werte möglichst gut mit den beobachteten Werten y übereinstimmen. Anders ausgedrückt, die Differenzen zwischen beobachteten und vorhergesagten Werten $y - \hat{y} = \varepsilon$, die Residuen, sollen möglichst klein sein. Daher scheint es zunächst naheliegend, die Regressionskoeffizienten so zu bestimmen, dass die, über alle Beobachtungseinheiten aufsummierten Residuen, also $\sum \varepsilon$, minimiert werden. Dieser Ansatz führt jedoch nicht zum gewünschten Ergebnis, da beliebig viele Mengen β_j existieren, bei denen die Summe der Residuen gleich null ist. Dies ist für alle diejenigen Mengen β_j der Fall, bei denen die vorhergesagten Werte durch den Schwerpunkt der beobachteten Verteilung, also durch den Punkt $(\bar{y},\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_k)$ gehen. Auf der Suche nach alternativen Verfahren entdeckten Carl F. Gauß (1795) und Adrien-Marie Legendre (1806) unabhängig voneinander, dass nicht die Summe der Residuen, sondern die Summe der quadrierten Residuen minimiert werden muss. Dieses Verfahren trägt daher den Namen Methode der kleinsten Quadrate (english: *method of least squares*, im Zusammenhang mit der linearen Regression meist auch *ordinary least squares* bzw. OLS genannt). Formal lautet die Minimierungsbedingung

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2\,, \qquad (7)$$

und die Schätzer für die Regressionskoeffizienten β_j lassen sich durch partielle Ableitung von Gleichung (7) nach β_j bestimmen. Dies resultiert in einem Gleichungssystem bei dem die Nullstelle das Minimum anzeigt.[3] Exemplarisch sei die Vorgehensweise zunächst für die partielle Ableitung nach β_1 etwas ausführlicher dargestellt. Bei der Ableitung von Gleichung (7) nach β_1 muss die Kettenregel – innere Ableitung mal äußere Ableitung – angewandt werden. Die äußere Ableitung von $\sum(\cdot)^2$ ist $2\sum(\cdot)$. Die innere Ableitung von $(y - \mathbf{x}\boldsymbol{\beta})$ nach β_1 beträgt $-x_{i1}$. Multipliziert man nun innere und äußere Ableitung und setzt das Ergebnis gleich null, ergibt sich

$$2 \sum_{i=1}^{n} -x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0$$

bzw.

$$-2 \sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0\,. \qquad (8)$$

Dieser Ausdruck lässt sich schließlich noch vereinfachen, indem beide Seiten der Gleichung durch -2 geteilt werden. Es ergibt sich somit

[3] Im Allgemeinen kann es sich bei den Nullstellen von Ableitungen um beide Formen von Extremwerten handeln, ein Minimum oder Maximum. Gleichung (7) beschreibt eine nach oben geöffnete Parabel, die nur über einen Extremwert, ein Minimum, verfügt.

$$\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) = 0\,. \tag{9}$$

Bildet man die partiellen Ableitungen nach allen zu bestimmenden Parametern β_j und setzt diese gleich null, so ergibt sich das folgende Gleichungssystem (vgl. Wooldridge 2009, S. 800 f.):

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}) &= 0\,.
\end{aligned} \tag{10}$$

Die erste Gleichung ergibt sich aus der partiellen ersten Ableitung nach β_0, die zweite aus der partiellen Ableitung nach β_1 usw. In Matrixnotation lässt sich dieses Gleichungssystem auch als

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \tag{11}$$

schreiben. Ausmultiplizieren und Umstellen ergibt

$$(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}\,. \tag{12}$$

Unter der Annahme, dass $(\mathbf{X}'\mathbf{X})$ vollen Rang hat, können wir beide Seiten von links mit der Inversen dieser Matrix, nämlich mit $(\mathbf{X}'\mathbf{X})^{-1}$, multiplizieren und erhalten

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,. \tag{13}$$

Diese Formel liefert die Schätzer für die Regressionskoeffizienten nach der Methode der kleinsten Quadrate. Der Vektor $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ löst also das Ausgangsproblem und minimiert die Summe der quadrierten Residuen $\sum(y - \hat{y})^2 = \sum \varepsilon^2$.

2.3 Annahmen der Kleinst-Quadrat-Methode

Die im letzten Abschnitt beschriebene Methode der kleinsten Quadrate ist an das Vorliegen bestimmter Voraussetzungen geknüpft (vgl. z. B. Berry 1993; Wooldridge 2009). Sind diese verletzt, sind die gewonnen Schätzer nicht mehr optimal. Zu den wichtigsten Voraussetzungen gehört:

– Die Variablen müssen metrisches Skalenniveau aufweisen, die unabhängigen Variablen dürfen auch als Dummy-Variablen kodierte kategoriale Merkmale enthalten.

- Die Daten müssen aus einer Zufallsstichprobe der interessierenden Population stammen. Dies gilt zumindest dann, wenn inferenzstatistische Schlüsse gezogen werden sollen (vgl. dazu Abschnitt 2.5). Soll lediglich das vorhandene Datenmaterial anhand eines Regressionsmodells beschrieben werden, ist diese Voraussetzung irrelevant.
- Die unabhängigen Variablen müssen ohne Messfehler gemessen sein.
- Ferner muss gelten, dass keine der unabhängigen Variablen sich als Linearkombination aus anderen unabhängigen Variablen bilden lässt und es sich bei keiner der unabhängigen Variaben um eine Kostante handelt. D. h. die Matrix $\mathbf{X}$ muss vollen Rang haben; es darf keine perfekte Multikollinearität vorliegen.
- Die Residuen müssen normalverteilt sein.
- Die Varianz der Residuen muss für jeden Wert der unabhängigen Variablen identisch sein; d. h. $\mathrm{Var}(\varepsilon|\mathbf{x}) = \mathrm{const}$; es muss also Homoskedastizität bestehen.
- Der Erwartungswert der Residuen muss für jede Kombination der unabhängigen Variablen null sein; d. h. $E(\varepsilon|\mathbf{x}) = 0$. Dies ist gleichbedeutend mit der Annahme, dass keine der unabhängigen Variablen mit dem Fehlerterm korreliert ist. In der ökonometrischen Literatur wird auch von strikter Exogenität gesprochen. Diese Voraussetzung bedingt auch, dass das Modell richtig spezifiziert sein muss. Es muss also einerseits alle bedeutsamen unabhängigen Variablen enthalten und darf keine für die Erklärung der abhängigen Variablen irrelevanten unabhängigen Variablen enthalten. Andererseits muss das Modell die richtige Parametrisierung aufweisen; die Prädiktoren müssen also in der gewählten Operationalisierung in einer linearen Beziehung zur untersuchten Variablen stehen.

Gelten diese Bedingungen, sind die nach der Methode der kleinsten Quadrate geschätzten Regressionskoeffizienten unverzerrt und weisen den kleinstmöglichen Standardfehler auf; sie sind also BLUE: best linear unbiased estimators. Im konkreten Fall einer empirischen Analyse werden die genannten Annahmen meist nur mehr oder weniger gut erfüllt sein. Dies hat zur Folge, dass die Schätzer für die Regressionskoeffizienten und/oder ihre Standardfehler vom Ideal eines effizienten und unverzerrten Schätzers abweichen. Um die Qualität von Regressionsanalysen einschätzen zu können, ist es wichtig zu wissen, welche Folgen die Verletzung der Annahmen hat.

Multikollinearität, Heteroskadastizität und nicht normalverteilte Residuen haben zur Folge, dass die Schätzer für die Standardfehler verzerrt sind. Verzerrte Standardfehler führen ihrerseits zu fehlerhaften Signifikanztests und fehlerhaften Konfidenzintervallen. Die Schätzer für die Regressionskoeffizienten dagegen bleiben von diesen Verletzungen der Annahmen unberührt, d. h. sie sind weiterhin unverzerrt.

Einen deutlich größeren Einfluss auf die Ergebnisse hat jedoch die Verletzung der verbleibenden Annahmen. Eine falsche Spezifikation des Modells führt zu verzerrten Schätzern für die Regressionskoeffizienten und die Standardfehler. Um verständlich zu machen, warum das so ist, betrachten wir folgenden Fall. Das wahre Modell sei

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \beta_m x_m + \varepsilon\,.$$

Nehmen wir jetzt an, ein Forscher wüßte nicht, dass x_m ein relevanter Einflussfaktor ist und spezifiziert daher das Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon^* ,$$

in dem x_m fehlt. Die Residuen des analysierten Modells entsprechen dann dem Fehlerterm des wahren Modells zuzüglich der, mit dem Regressionskoeffizienten gewichteten, nicht inkludierten Variablen x_m; also: $\varepsilon^* = \beta_m x_m + \varepsilon$. Ist x_m mit mindestens einer der anderen unabhängigen Variablen x_1 bis x_k korreliert – und das wird nahezu immer der Fall sein – sind die Residuen im analysierten Modell mit den unabhängigen Variablen korreliert. Warum dies so ist, wird aus Gleichung (13) deutlich. Die Korrelation zwischen den unabhängigen Variablen hat einen Einfluss auf die Berechnung der Regressionskoeffizienten. Entsprechend führt eine Berechnung dieser Koeffizienten unter Ausschluss von Merkmalen, die sowohl mit der abhängigen Variablen als auch mit den unabhängigen Variablen korreliert sind, zu verzerrten Schätzungen – dem sog. *omitted variable bias*. Die einzige Möglichkeit, der Gefahr fehlspezifizierter Modelle zu begegnen, besteht in einer sorgfältigen theoretischen Fundierung der Modelle und einer adäquaten Operationalisierung der theoretischen Begriffe (vgl. für ein entsprechendes Beispiel etwa Best 2009).

Ein weiteres, weitverbreitetes Problem sind Messfehler in den unabhängigen Variablen. Einerlei, ob es sich um zufällige oder systematische Messfehler handelt, führen nicht vollständig reliabel gemessene Variablen zu verzerrten Schätzungen der Regressionskoeffizienten und ihrer Standardfehler (Cohen et al. 2003, S. 119).[4] Hier hilft nur, die Messungen durch bessere Erhebungsinstrumente und die Verwendung geeigneter Skalierungsverfahren zu verbessern. Liegen für die interessierenden Merkmale jeweils mehrere Indikatoren vor, bietet sich der Einsatz von Strukturgleichungsmodellen an, die entsprechende Messfehler in der Modellierung explizit berücksichtigen (vgl. Kapitel 29 in diesem Handbuch).

Damit soll dieser kurze Abschnitt zu den Anwendungsvoraussetzungen der linearen Regression beendet werden. Eine ausführlichere Diskussion dieser Annahmen sowie der Verfahren zu ihrer Überprüfung bietet Kapitel 25 in diesem Handbuch.

2.4 Die Bestimmung der Modellgüte

Nach der Methode der kleinsten Quadrate lassen sich für jede beliebige Kombination aus abhängiger und unabhängigen Variablen Schätzer für β_j gewinnen, die für die jeweils betrachtete Menge von Variablen die kleinste Summe der quadrierten Fehler liefert, also die bestmögliche Anpassung von beobachteten und erwarteten Werten gewährleistet. „Bestmögliche“ Anpassung bedeutet jedoch nicht, dass jedes Regressionsmodell denselben Grad an Anpassung an die Daten aufweist. In manchen Fällen wird die Anpassung höher sein, in anderen geringer. Für jedes Regressionsmodell, das bestimmt wurde, stellt sich daher die Frage, wie gut seine Anpassung an die Daten ist. Die Antwort auf diese Frage wird davon abhängen, wie groß die Diskrepanz zwischen unter dem Modell erwarteten Werten ($\hat{y}$) und den beobachteten Werten (y) ist. Das Modell ist umso besser, je besser es die beobachteten Unterschiede der

[4] Man könnte annehmen, dass zufällige Messfehler der unabhängigen Variablen zu einer Unterschätzung der Regressionskoeffizienten führen. Dies ist jedoch leider nicht immer der Fall (Cohen et al. 2003).

Untersuchungseinheiten in Bezug auf y reproduzieren kann. Zur Operationalisierung dieser Vorstellung greift man auf die Varianz der abhängigen Variablen zurück: je höher der Anteil dieser Varianz ist, den das Modell „erklären“ kann, desto besser ist es. Diese Vorstellung wird in der Maßzahl

$$R^2 = \frac{\text{Erklärte Streuung}}{\text{Gesamte Streuung}} = \frac{SSR}{SST} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} \tag{14}$$

zum Ausdruck gebracht.[5] Hierbei steht SSR für die durch die Regression erklärte Streuung (sum of squares due to regression) und SST für die Gesamtstreuung der Variablen (sum of squares total). Die Maßzahl R^2 kann zwischen 0 und 1 variieren und wird als Anteil erklärter Varianz, teils auch als Determinationskoeffizient oder Bestimmtheitsmaß bezeichnet. Je höher ihr Wert, desto größer der Anteil, der durch das Regressionsmodell erklärten Varianz, d. h. desto besser die Anpassung des Modells an die Daten.

Die Verwendung von R^2 ist nicht ganz unumstritten (vgl. zusammenfassend Urban & Mayerl 2006, S. 59 ff. und 109 ff.). Ein Problem besteht darin, dass diese Maßzahl mit jeder zusätzlich in das Modell aufgenommenen Variablen steigt, auch wenn die zusätzliche Variable nicht wesentlich zur Verbesserung des Modells beiträgt. Aus diesem Grunde können Modelle mit unterschiedlich vielen unabhängigen Variablen auch nicht zuverlässig über R^2 miteinander verglichen werden. Ein weiteres Problem von R^2 ist, dass sein Erwartungswert auch, wenn kein Zusammenhang zwischen x_j und y besteht, nicht null ist. Eine Lösung dieser beiden Probleme stellt die Verwendung des sog. korrigierten R^2 (englisch: adjusted R^2) dar. Diese Maßzahl ist als

$$R^2_{\text{korr}} = 1 - \frac{n-1}{n-k-1}(1 - R^2) \tag{15}$$

definiert. Während R^2 bei der Hinzunahme weiterer Variablen nur steigen kann, kann R^2_{korr} auch kleiner werden, wenn die zusätzliche Variable das Modell nicht verbessert. Somit bestraft R^2_{korr} die Hinzunahme „irrelevanter“ Variablen. Sind die abhängige und die unabhängigen Variablen insgesamt nicht miteinander korreliert, kann R^2_{korr} sogar negativ werden.

Ein weiteres Problem von R^2, das auch für R^2_{korr} gilt, besteht darin, dass es nicht nur von der erklärten Varianz der abhängigen Variablen abhängt, sondern auch von der Varianz der Prädiktoren. Damit ist ein Vergleich von Regressionsmodellen aus verschiedenen Populationen, in denen sich diese Faktoren in verschiedenem Ausmaß unterscheiden, problematisch. Daher sollten entsprechende Vergleiche mit Vorsicht erfolgen und auch die Unterschiede in den Varianzen und Regressionskoeffizienten berücksichtigen. Unseres Erachtens bleiben diese Maßzahlen trotz der genannten Schwierigkeiten nützliche Werkzeuge zur Beschreibung eines Regressionsmodells. Insbesondere R^2_{korr} kann unseres Erachtens bei der Entscheidung zwischen verschiedenen Modellen für dieselbe abhängige Variable nützlich sein.

Ungeachtet der genannten Schwierigkeiten muss bei der Interpretation von R^2 bzw. R^2_{korr} zudem berücksichtigt werden, dass die lineare Regression nur *lineare* Zusammenhänge abbilden kann und diese Maßzahlen folglich nur die Stärke des linearen

[5] In der bivariaten Regression entspricht R^2 der quadrierten Korrelation zwischen x und y.

Zusammenhangs zwischen y und den x_j widerspiegeln. Ist der lineare Zusammenhang klein oder gar gleich null, kann dennoch ein anderer, nichtlinearer Zusammenhang zwischen den analysierten Merkmalen bestehen. Dies lässt sich z. B. mit grafischen Verfahren klären (vgl. Kapitel 34 in diesem Handbuch). Liegt ein nichtlinearer Zusammenhang vor, kann dieser unter Umständen dennoch im Rahmen der linearen Regression modelliert werden, indem eine alternative Parametrisierung für die beteiligten Variablen gewählt wird (siehe Kapitel 26 in diesem Handbuch).

2.5 Die statistische Absicherung der Regressionsergebnisse

In der Regel werden die Koeffizienten einer Regressionsanalyse auf der Basis von Stichprobendaten geschätzt. In dieser Situation stellt sich die Frage, ob die entsprechenden Ergebnisse auch für die Grundgesamtheit gelten, aus der die Stichprobe stammt. Handelt es sich bei der Stichprobe um eine Zufallsstichprobe, kann diese Frage mithilfe der Inferenzstatistik beantwortet werden. Die im Folgenden vorgestellten Verfahren gehen von der vereinfachenden Annahme aus, dass die Daten aus einer einfachen Zufallsstichprobe stammen. Entsprechende Aussagen lassen sich im Prinzip mit denselben Verfahren auch für Daten aus mehrstufigen und/oder geschichteten Zufallsstichproben gewinnen, allerdings sind die entsprechenden Formeln komplizierter. Daher sei für diesen Fall auf die einschlägige Literatur verwiesen (z. B. Bacher 2009; Lee & Forthofer 2006).

Werden die Regressionskoeffizienten auf der Basis von Stichprobendaten geschätzt, wird dies meist durch das Hinzufügen eines Zirkumflex gekennzeichnet. Das Grundmodell der Regressionsgleichung wird dann zu

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \varepsilon_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ij} + \varepsilon_i$$

oder in Matrixnotation

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}\,. \tag{6'}$$

Entsprechend wird aus Gleichung (13) zur Bestimmung der Regressionskoeffizienten

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,. \tag{13'}$$

Liegen die Schätzungen $\hat{\beta}_j$ vor, stellt sich die Frage, ob die Ergebnisse der Regressionsanalyse mit hinreichender Sicherheit Aussagen über die Grundgesamtheit erlauben. Es muss also untersucht werden, ob die Effekte der unabhängigen Variablen auf das abhängige Merkmal statistisch signifikant sind. Dazu dient der folgende statistische Test, mit dem geprüft werden kann, ob ein Regressionskoeffizient statistisch signifikant von einem gegebenen Wert a abweicht. Die entsprechenden zweiseitigen statistischen Hypothesen[6] zu diesem Test lauten

[6] Eine ausführliche Einführung in die Logik des statistischen Testens bietet Kapitel 8 in diesem Handbuch.

$$H_0\colon \beta_j = a$$
$$H_1\colon \beta_j \neq a$$

und können anhand der statistischen Prüfgröße

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a}{s_{\hat{\beta}_j}} \tag{16}$$

getestet werden.[7] Unter den üblichen OLS-Annahmen folgt diese Prüfgröße einer t-Verteilung mit $n-k-1$ Freiheitsgraden. Bei einer hinreichend großen Stichprobengröße sind die $\hat{\beta}_j$ normalverteilt und die präsentierte Prüfgröße geht in eine Standardnormalverteilung über. Auf Basis dieser Prüfgröße lassen sich nun beliebige statistische Hypothesen über die Differenz von $\hat{\beta}_j$ und dem interessierenden Wert a prüfen. In der gängigen Standardsoftware wird typischerweise der zweiseitige Test für $a = 0$ ausgegeben. Das entsprechende Hypothesenpaar lautet dann entsprechend

$$H_0\colon \beta_j = 0$$
$$H_1\colon \beta_j \neq 0\,.$$

Die untersuchte Frage lautet also, ob der auf der Basis von Stichprobendaten geschätzte Wert $\hat{\beta}_j$ mit einer gegebenen Sicherheit in der Grundgesamt von 0 verschieden ist. Es wird demnach gefragt, ob davon ausgegangen werden kann, dass das Merkmal x_j auch in der Grundgesamtheit einen Einfluss auf das untersuchte abhängige Merkmal hat. Über das Ausmaß der Sicherheit, mit dem eine solche Aussage getroffen werden kann, entscheidet das Signifikanzniveau, welches typischerweise bei einer Irrtumswahrscheinlichkeit von 0,01 oder 0,05 festgelegt wird.

Neben den Hypothesen zum Vergleich eines Regressionskoeffizienten mit einem Referenzwert lassen sich auch Hypothesen über die Gleichheit bzw. Ungleichheit zweier Regressionskoeffizienten desselben Modells prüfen. Nehmen wir an, in einem Modell würde der Einfluss der Lebensweise und der Einfluss der genetischen Disposition auf die Lebenserwartung untersucht. Eine naheliegende Frage ist dann, ob der Einfluss der genetischen Disposition (β_1) auf die Lebenserwartung größer ist als der Einfluss der Lebensweise (β_2). Die einseitigen Hypothesen lauten:

$$H_0\colon \beta_1 \leq \beta_2$$
$$H_1\colon \beta_1 > \beta_2\,.$$

Die entsprechende Testgröße ist wiederum t-verteilt und hat die Form

[7] Der Standardfehler der $\hat{\beta}_j$ ergibt sich aus

$$s_{\hat{\beta}_j} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,/\, (n-k-1)}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \left(1 - R_j^2\right)}}\,,$$

wobei R_j^2 für den Anteil erklärter Varianz von x_j steht, der durch die anderen unabhängigen Variablen aufgeklärt wird (vgl. Wooldridge 2009, S. 89).

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{s^2_{\hat{\beta}_1} + s^2_{\hat{\beta}_2} - 2s_{\hat{\beta}_1\hat{\beta}_2}}}\,, \tag{17}$$

wobei $s_{\hat{\beta}_1\hat{\beta}_2}$ die Kovarianz zwischen den Schätzern für $\hat{\beta}_1$ und $\hat{\beta}_2$ bezeichnet. Schätzer für die Varianzen und Kovarianz der Regressionskoeffizienten werden von gängigen Statistikprogrammen bereitgestellt.[8]

Es können jedoch nicht nur einzelne Koeffizienten, sondern auch das gesamte Modell auf seine Erklärungskraft hin überprüft werden. Die entsprechenden statistischen Hypothesen lauten in diesem Fall

$$\begin{aligned} &H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1\text{: } \beta_j \neq 0 \text{ für mindestens ein } j\,. \end{aligned} \tag{18}$$

Dieser Test erinnert an den globalen Test bei der Varianzanalyse (vgl. Kapitel 19 in diesem Handbuch) und wie dort ist die Prüfgröße auch hier F-verteilt, wobei sich F als

$$F = \frac{\sum(\hat{y} - \bar{y})^2/k}{\sum(y - \hat{y})^2/(n-k-1)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} \tag{19}$$

ergibt.[9] Im Zähler der Prüfgröße steht die mittlere durch die Regression erklärte Streuung MSR (*mean square regression*). Im Nenner steht die mittlere nicht erklärte Streuung MSE (*mean square error*). Eine alternative Definition derselben Prüfgröße lautet

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

Liegt der empirisch ermittelte Wert der Prüfgröße über einem zuvor festgelegten kritischen F-Wert mit $df_1 = k-1$ und $df_2 = n-k-1$, dann wird H_0 verworfen.

In manchen Fällen wird das Interesse weniger einem globalen Test für alle Koeffizienten eines Modells gelten als vielmehr dem Vergleich zweier verschiedener Modelle. Gehen wir von einem Regressionsmodell mit k unabhängigen Variablen x_j aus. Eine Frage könnte sein, ob eine Untermenge von d, $d < k$, dieser Variablen einen Beitrag zur Erklärung der abhängigen Variablen leistet. Zur Vereinfachung der Notation soll angenommen werden, dass die d interessierenden Variablen in der Regressionsgleichung die ersten sind. Die beiden Modelle können dann wie folgt geschrieben werden:

$$\begin{aligned} &\text{Modell 1: } y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_d x_d + \hat{\beta}_{d+1} x_{d+1} + \cdots + \hat{\beta}_k x_k + \varepsilon \\ &\text{Modell 2: } y = \hat{\beta}_0 + \hat{\beta}_{d+1} x_{d+1} + \cdots + \hat{\beta}_k x_k + \varepsilon\,. \end{aligned}$$

[8] Beispielsweise in SPSS, indem auf dem Unterkommando `/STATISTICS` der Regressionsprozedur das Schlüsselwort `BCOV` angegeben wird.

[9] Hier und im Folgenden steht SSR (sum of squares due to regression) für die durch die Regression erklärte Streuung $\sum(\hat{y} - \bar{y})^2$; SSE (sum of squared errors) steht für die nicht erklärte Streuung $\sum(y - \hat{y})^2$. MSR und MSE sind entsprechend die mittlere erklärte Streuung bzw. die mittlere nicht erklärte Streuung. Zu dieser Schreibweise vgl. auch den Exkurs zu mittleren Quadratsummen in Kapitel 19 in diesem Handbuch.

Modell 2 ist in Modell 1 geschachtelt (englisch: *nested*), weil das Modell an derselben Stichprobe untersucht wird und die in ihm enthaltenen Parameter $\hat{\beta}_{d+1}, \ldots, \hat{\beta}_k$ eine Untermenge der in Modell 1 enthaltenen Parameter ist. Weil die Koeffizienten $\hat{\beta}_1$ bis $\hat{\beta}_d$ in Modell 2 auf null gesetzt sind, wird dieses Modell auch als restriktives, Modell 1 als nicht oder weniger restriktives Modell bezeichnet.

Die Vermutung, dass die ersten d Koeffizienten ohne Bedeutung für y sind, lässt sich in die statistische Hypothesen

$$\begin{aligned} &H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_d = 0 \\ &H_1\text{: } \beta_j \neq 0 \text{ für mindestens ein } j \leq d \end{aligned} \tag{20}$$

übersetzen. Die Prüfgröße zur Beurteilung dieser Nullhypothese ist wiederum F-verteilt und lautet

$$F = \frac{\left(\sum (y - \hat{y}_{\mathrm{r}})^2 - \sum (y - \hat{y}_{\mathrm{nr}})^2\right)/d}{\sum (y - \hat{y}_{\mathrm{nr}})^2/(n-k-1)} = \frac{(SSE_{\mathrm{r}} - SSE_{\mathrm{nr}})/d}{SSE_{\mathrm{nr}}/(n-k-1)} = \frac{MSE_{\mathrm{r}} - MSE_{\mathrm{nr}}}{MSE_{\mathrm{nr}}}\,. \tag{21}$$

Die Kenngrößen des restriktiven Modells, hier Modell 2, sind mit r bezeichnet, die des weniger restriktiven Modells mit nr. Die Beurteilung der Hypothesen erfolgt wieder, indem der empirisch ermittelte F-Wert mit einem dem gewählten Signifikanzniveau entsprechenden kritischen F-Wert mit $df_1 = d$ und $df_2 = n - k - 1$ verglichen wird.

Der in Gleichung (21) genannte Test ist besonders nützlich, wenn geprüft werden soll, ob eine kategoriale Variable – z. B. der Familienstand –, die durch mehrere Dummy-Variablen repräsentiert wird, einen statistisch signifikanten Einfluss auf die abhängige Variable hat. Der in Gleichung (16) aufgeführte t-Test hilft in diesem Fall nicht weiter, weil er nur die Überprüfung jeweils *eines* Regressionskoeffizienten erlaubt. Im Falle einer kategorialen Variablen mit m Kategorien liegen jedoch $m - 1$ Regressionskoeffizienten vor und damit muss der in Gleichung (21) wiedergegebene Test verwendet werden. Die entsprechenden Modelle lauten

$$\begin{aligned} &\text{Modell 1: } y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{D_1} D_1 + \hat{\beta}_{D_2} D_2 + \cdots + \hat{\beta}_{D_{m-1}} D_{m-1} + \varepsilon \\ &\text{Modell 2: } y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \varepsilon\,. \end{aligned}$$

Wiederum ist Modell 2 in Modell 1 geschachtelt, Modell 2 ist damit restriktiver als Modell 1. Als statistische Hypothesen ergeben sich nun

$$\begin{aligned} &H_0\text{: } \beta_{D_1} = \beta_{D_2} = \cdots = \beta_{D_{m-1}} = 0 \\ &H_1\text{: } \beta_{D_j} \neq 0 \text{ für mindestens ein } D_j\,. \end{aligned}$$

Nach Wahl des Signifikanzniveaus wird man nun die beiden Modelle schätzen, für beide die Summe der quartierten Fehler (SSE) ermitteln und damit die empirische Prüfgröße nach Gleichung (21) berechnen.

Ein anderer Testfall liegt vor, wenn es um die Frage geht, ob ein gegebenes Regressionsmodell in zwei verschiedenen Populationen zu unterschiedlichen Ergebnissen führt. Nehmen wir an, es soll geprüft werden, ob eine bestimmte Einkommensgleichung gleichermaßen für Männer und Frauen, für Deutschland und England oder für Daten

aus einem Jahr t_0 und einem Jahr t_1 gilt. Wir haben es jeweils mit zwei Modellen folgender Form zu tun:

$$\text{Modell 1: } y_1 = \hat{\beta}_{10} + \hat{\beta}_{11}x_1 + \hat{\beta}_{12}x_2 + \cdots + \hat{\beta}_{1k}x_k + \varepsilon$$
$$\text{Modell 2: } y_1 = \hat{\beta}_{20} + \hat{\beta}_{21}x_1 + \hat{\beta}_{22}x_2 + \cdots + \hat{\beta}_{2k}x_k + \varepsilon\,.$$

Jedes Modell enthält dieselben Variablen, die jedoch aus unterschiedlichen, voneinander unabhängigen Stichproben stammen. Der Effekt von x_1 in der ersten Stichprobe wird entsprechend mit β_{11} bezeichnet, der Effekt derselben Variablen in der zweiten Gruppe mit β_{21} usw. Die interessierenden Hypothesen lauten entsprechend

$$H_0\text{: } \beta_{1j} = \beta_{2j} \text{ für alle } j = 1, \ldots, k$$
$$H_1\text{: } \beta_{1j} \neq \beta_{2j} \text{ für mindestens ein } j\,.$$

Da die Modelle in unterschiedlichen Stichproben geschätzt werden, sind sie nicht geschachtelt und daher kann der unter Gleichung (21) angegebene Test nicht verwendet werden. Für diesen Fall steht der sog. Chow-Test zur Verfügung. Bei diesem Test handelt es sich wiederum um eine F-verteilte Größe, die als

$$F = \frac{\left(SSE_p - (SSE_1 + SSE_2)\right)/(k+1)}{(SSE_1 + SSE_2)/\left(n - 2(k+1)\right)} \tag{22}$$

definiert ist. Bei SSE handelt es sich abermals um die Summe der quadrierten Fehler. Wie die Indizes anzeigen, müssen die SSE-Werte aus drei verschiedenen Regressionen verwendet werden: SSE_p stammt aus einer gemeinsamen (gepoolten) Regression, SSE_1 stammt aus einer Regression in der ersten Gruppe und SSE_2 aus einer Regression in der zweiten Gruppe. Zur Feststellung statistischer Signifikanz wird das Resultat der Prüfgröße wiederum mit dem entsprechenden kritischen F-Wert mit $df_1 = k + 1$ und $df_2 = n - 2(k + 1)$ verglichen.

2.6 Die Interpretation der Regressionskoeffizienten

Ist das Regressionsmodell statistisch abgesichert, stellt sich immer noch die Frage nach der inhaltlichen Bedeutung und der substanziellen Signifikanz der Ergebnisse. Betrachten wir zunächst die unstandardisierten Regressionskoeffizienten β_j, die auch als Effektgrößen oder Effektstärken bezeichnet werden (zum Problem standardisierter Koeffizienten vgl. den nächsten Abschnitt). Die häufig verwendete Interpretation der Regressionskoeffizienten, nach der eine Erhöhung von x_j um eine Einheit zu einer Veränderung von y um β_j Einheiten führt, ist streng genommen in den meisten Fällen falsch. Insbesondere wenn Daten aus einer Querschnittserhebung verwendet werden, ist eine derartige Interpretation, bei der es sich um eine Vorhersage handelt, nicht zulässig (für die Panelregression vgl. Kapitel 36 in diesem Handbuch). Richtig ist vielmehr, dass sich der Erwartungswert von y bei Analyseeinheiten, deren Wert für x_j um eine Einheit größer ist als bei anderen Analyseeinheiten, um β_j Einheiten unterscheidet.

Nehmen wir an, wir hätten ein einfaches lineares Regressionsmodell, um zu untersuchen, welchen Einfluss die Körpergröße auf das Körpergewicht hat. Wenn wir

Körpergewicht in Kilogramm und Körpergröße in Zentimeter gemessen haben, bedeutet ein Steigungskoeffizient von 0,7, dass von zwei Personen, deren Körpergröße sich um einen Zentimeter unterscheidet, die größere im Durchschnitt 700 Gramm mehr wiegt. Bei dieser Aussage handelt es sich um eine Schätzung, genauer eine Punktschätzung auf Basis von Stichprobendaten, die mit Unsicherheit behaftet ist. Es ist daher ratsam, auch die Konfidenzintervalle der Regressionskoeffizienten zu analysieren (eine Einführung in Konfidenzintervalle liefert Kapitel 8 in diesem Handbuch). Nehmen wir an, das 95 %-Konfidenzintervall für den Effekt der Körpergröße auf das Gewicht hätte die Grenzen [0,55 ; 0,85]. Dann könnten wir mit einer Wahrscheinlichkeit von 95 % davon ausgehen, dass das Intervall von 550 bis 850 Gramm den wahren Gewichtsunterschied, der mit einem Größenunterschied von einem Zentimeter einhergeht, einschließt.

Die Interpretation der Regressionskoeffizienten für mit den Werten 0 und 1 kodierte Dummy-Variablen folgt derselben Logik. Nehmen wir an, wir hätten im soeben genannten Modell für das Körpergewicht auch das Geschlecht aufgenommen und zwar mit der Kodierung 0 für weiblich und 1 für männlich. Ein Regressionskoeffizient von 6,0 würde bedeuten, dass Männer durchschnittlich sechs Kilogramm mehr wiegen als Frauen und zwar bei *gleicher* Größe. Gerade die letzte Aussage verweist auf eine große Stärke der Regressionsanalyse sowie multivariater Verfahren im Allgemeinen. In der Regressionsanalyse wird der Effekt einer Variablen *unter Konstanthaltung* aller anderen Variablen im Modell geschätzt. Da es in den Sozialwissenschaften häufig nicht möglich ist, Daten unter experimentellen Bedingungen zu generieren, die konstante Bedingungen garantieren würden, ist es umso wichtiger, dass das „Konstanthalten“ von „Störfaktoren“ ex post durch die Datenanalyse erfolgt. Dadurch erklärt sich die große Attraktivität und Bedeutung multivariater Verfahren im Allgemeinen und Verfahren der Regressionsanalyse im Besonderen (vgl. auch Kapitel 2 in diesem Handbuch).

Die Interpretation der Regressionskoeffizienten wird oftmals durch eine veränderte Skalierung der unabhängigen Variablen erleichtert. Nehmen wir an, in einem Modell zur Erklärung der Lebenserwartung in Jahren finden wir einen Effekt des Einkommens in Euro von 0,0001. Die Lebenserwartung steigt also um 0,0001 Jahre je zusätzlich verdientem Euro. Wird das Einkommen dagegen in 10.000 € gemessen, verändert sich der Koeffizient zu 1; ein Wert, der besser interpretier- und besser kommunizierbar ist: die Lebenserwartung von Personen, deren Einkommen sich um 10.000 € unterscheidet, wird sich durchschnittlich um 1 Jahr unterscheiden.

Bisher haben wir lediglich die Steigungskoeffizienten betrachtet und den Achsenabschnitt β_0 vernachlässigt. Dieser gibt den Erwartungswert von y für den Fall an, dass alle x_j null sind. In den meisten Analysen handelt es sich dabei um einen unter inhaltlichen Gesichtspunkten vollkommen uninteressanten, oftmals auch unsinnigen Wert. Nehmen wir noch einmal als Beispiel das Modell zur Erklärung des Körpergewichts in Kilogramm mit den Prädiktoren Körpergröße in Zentimetern und Geschlecht in der oben genannten Kodierung. Nehmen wir ferner an, die Daten stammen von Erwachsenen und der Wertebereich der Körpergröße betrage in der Stichprobe 150 bis 200 cm. Das Ergebnis der Analyse sei

$$\widehat{\text{Körpergewicht}} = -50 + 0{,}7 \cdot \text{Körpergröße} \; + 6 \cdot \text{Mann}\,.$$

Gemäß dieser Gleichung sollte eine Null Zentimeter große Frau −50 kg wiegen. Dieser Wert ist aus mehreren Gründen unsinnig. Erstens gibt es eine solche Frau nicht; zweitens, selbst wenn es eine solche Frau gäbe, in unseren Daten haben wir sie nicht beobachtet. Die kleinste in unseren Daten vorhandene Person ist 150 cm groß. Daher sollten auf Basis dieser Untersuchung keine Aussagen über Personen gemacht werden, die kleiner als 150 cm sind. Doch zurück zum Achsenabschnitt. Dieser kann sinnvoll interpretiert werden, wenn die Variablen vor der Analyse zentriert werden. Die Zentrierung erfolgt meist auf den Mittelwert. Es kann jedoch sinnvoll sein auf andere Werte, die die Interpretation des Achsenabschnitts verbessern, zu zentrieren. Nehmen wir an, von der Körpergröße würde die durchschnittliche Größe der Frauen – 166 cm – abgezogen. Mit der entsprechend reskalierten Variable ergäbe sich dann

$$\widehat{\text{Körpergewicht}} = 66 + 0{,}7 \cdot \text{Körpergröße}_C + 6 \cdot \text{Mann}$$

als Regressionsgleichung. Jetzt wäre der Achsenabschnitt zu interpretieren als Erwartungswert des Gewichts von Frauen durchschnittlicher Körpergröße; eine durchaus interessante Information, die sinnvoll interpretiert werden kann.

2.7 Standardisierte Regressionskoeffizienten und ihre Probleme

Die bisher zur Interpretation herangezogenen Koeffizienten geben Auskunft über die absolute Größe von Effekten. Ein typisches Problem sozialwissenschaftlicher Anwendungen der linearen Regression besteht jedoch darin, dass die Einheiten der verwendeten Merkmale oft beliebig und zudem von Merkmal zu Merkmal verschieden sind. Um dennoch etwas über die *relative Bedeutung* der verschiedenen Merkmale sagen zu können, werden diese oft „standardisiert“, also auf eine „gemeinsame“ Skala gebracht. Dies geschieht typischerweise, indem der Steigungskoeffizient mit der Standardabweichung der unabhängigen Variablen multipliziert und durch die Standardabweichung der abhängigen Variablen dividiert wird:

$$B_j^* = \beta_j \frac{\sigma_{x_j}}{\sigma_y} \,. \tag{23}$$

Die standardisierten Koeffizienten geben an, um welchen Teil einer Standardabweichung sich der Erwartungswert von y unterscheidet, wenn zwei Einheiten verglichen werden, die auf der unabhängigen Variablen eine Standardabweichung auseinander liegen. Die Standardisierung erfolgt somit, indem die untersuchten Merkmale jeweils auf ihre Standardabweichung als neue, gemeinsame Einheit bezogen werden. In den Sozialwissenschaften ist es gängige Praxis, die standardisierten Koeffizienten, häufig ausschließlich diese, zu berichten und zu interpretieren. Dabei wird davon ausgegangen, dass der relative Einfluss eines Prädiktors auf die untersuchte abhängige Variable um so größer ist, je höher der Betrag ihres standardisierten Regressionskoeffizienten ist.

Die Verwendung standardisierter Regressionskoeffizienten wurde aus verschiedenen Gründen kritisiert (vgl. Bring 1994; Urban & Mayerl 2006, S. 103 ff.). So wurde darauf hingewiesen, dass in die B_j^* zwei Konzepte eingehen: die Effektstärke und die Streuung der Variablen. Diese Sachverhalte sollten jedoch besser getrennt untersucht und interpretiert werden. Eine weitere Kritik lautet, dass die standardisierten Koeffizienten von

den Eigenschaften der jeweiligen Stichprobe abhängen, also von den jeweils beobachteten Standardabweichungen der unabhängigen und abhängigen Variablen sowie der Beziehung zwischen unabhängiger und abhängiger Variable. Häufig wird sich jedoch die Standardabweichung eines Merkmals zwischen zwei Stichproben unterscheiden. So könnte beispielsweise die Streuung der Einkommen von männlichen und weiblichen Beschäftigten verschieden sein. Entsprechend können die standardisierten Koeffizienten eines in zwei verschiedenen Populationen geschätzten Modells nicht ohne weiteres miteinander verglichen werden.

Doch auch der Vergleich der standardisierten Koeffizienten *innerhalb* eines Modells kann problematisch sein. Dies soll anhand des nachfolgenden Beispiels verdeutlicht werden (vgl. Bring 1994, S. 211). Nehmen wir an,

$$\widehat{\text{Einkommen}} = \beta_0 + \beta_1 \,\text{Berufserfahrung} + \beta_2 \,\text{Ausbildungsjahre}$$

sei das uns interessierende Modell. β_1 gibt dabei den Effekt der Berufserfahrung auf das Einkommen *unter Konstanthaltung* der Ausbildungsdauer wieder. Der standardisierte Effekt für die Berufserfahrung berechnet sich wie oben angegeben, indem β_1 mit der Standardabweichung der Berufserfahrung σ_{x_1} multipliziert wird. Diese Vorgehensweise ist laut Bring (1994) inkonsistent, weil sich β_1 auf einen konditionalen Sachverhalt *(unter Konstanthaltung von x_j)* bezieht, während σ_{x_1} ein Parameter der gesamten Population ist. Das Problem besteht demnach darin, dass sich der Steigungskoeffizient und die Standardabweichung, die beide in die Berechnung der standardisierten Koeffizienten eingehen, auf unterschiedliche Populationen beziehen. Als Ausweg schlägt Bring vor, statt der einfachen Standardabweichung die partielle Standardabweichung, also letztlich die über die Gruppen der anderen unabhängigen Variablen hinweg gemittelte Standardabweichung von x_j zu verwenden.

Eine weitere Kritik lautet, dass die standardisierten Regressionskoeffizienten nicht notwendigerweise den Beitrag der unabhängigen Variablen zur erklärten Varianz widerspiegeln. Die Interpretation der standardisierten Koeffizienten, nach der das Merkmal mit dem betragsmäßig höchsten Koeffizienten am stärksten zur erklärten Varianz beiträgt, das Merkmal mit dem betragsmäßig nächst höchsten Koeffizienten den zweitgrößten Beitrag zur erklärten Varianz leistet etc., ist nicht immer richtig. Die standardisierten Koeffizienten reflektieren nicht notwendigerweise, welches Merkmal am meisten zu R^2 beiträgt. Dies gilt nur für den Fall, dass die unabhängigen Variablen unkorreliert sind. Dann entspricht R^2 der Summe der quadrierten Korrelationskoeffizienten zwischen jeweils einer unabhängigen und der abhängigen Variablen. Da in diesem Fall die Korrelation dem standardisierten Regressionskoeffizient entspricht, entspricht R^2 der Summe der quadrierten standardisierten Regressionskoeffizienten. Bei unkorrelierten unabhängigen Variablen lässt sich R^2 also eindeutig und vollständig in die Beiträge der einzelnen unabhängigen Variablen zerlegen.[10] Die Variable mit dem größten Einfluss ist dann diejenige, welche am meisten zu R^2 beiträgt. Oder anders ausgedrückt: wird die Variable mit dem größten standardisierten Koeffizienten aus der

[10] Ein Sachverhalt, der in der Forschung praktisch nie vorkommt. Im Übrigen bräuchte man in dieser Situation ohnehin keine multivariaten Modelle. Die relative Einflussstärke eines Merkmals kann dann auch durch eine bivariate Analyse ermittelt werden.

Gleichung ausgeschlossen, sinkt R^2 mehr als beim Ausschluss jeder anderen Variablen. Für den üblicherweise vorliegenden Fall korrelierter Prädiktoren ist die Zerlegung der erklärten Varianz komplizierter, dann gilt

$$R^2 = \sum_{j=1}^{p} B_j^{*2} + 2 \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} B_j^* B_k^* \rho_{jk} \tag{24}$$

mit B_j^*, B_k^* als standardisierte Regressionskoeffizienten und ρ_{jk} als Korrelation zwischen x_j und x_k (vgl. Grömping 2007, S. 140).

Bring (1994) hat vorgeschlagen, die relative Bedeutung der einzelnen Prädiktoren durch das Produkt der Korrelation zwischen unabhängiger und abhängiger Variablen mit dem entsprechenden (unstandardisierten) Regressionskoeffizienten zu erfassen. Diese Maßzahl hat den Vorteil, dass sie sich über alle unabhängigen Variablen hinweg zu R^2 aufsummiert. Es gilt also

$$R^2 = \sum_{j=1}^{k} \beta_j \rho_{jy} \,. \tag{25}$$

Die relative Bedeutung einer Variablen würde also durch das Produkt des unstandardisierten Koeffizienten β_j mit der Korrelation ρ_{jy} bestimmt werden. Das Problem dieses einfachen Maßes ist allerdings, dass es negativ wird, wenn β_j und ρ_{jy} unterschiedliche Vorzeichen haben.

Mittlerweile liegt eine Reihe von Vorschlägen zu alternativen Maßzahlen vor, die die Beschränkungen und Probleme der standardisierten Koeffizienten zu vermeiden suchen und die relative Bedeutung von Merkmalen konsistent messen sollen. In einem neueren Beitrag vergleichen Chao et al. (2008) sechs Ansätze zur Bestimmung der relativen Bedeutung von unabhängigen Variablen. Zunächst untersuchen sie, ob die vorgeschlagenen Koeffizienten folgende Kriterien erfüllen: (a) die Koeffizienten der relativen Bedeutung sollen sich zu R^2 summieren, (b) keiner dieser Koeffizienten soll negativ sein, (c) das Ergebnis muss unabhängig von der Reihenfolge sein, in der die unabhängigen Variablen ins Modell aufgenommen werden. Nur zwei der sechs untersuchten Maßzahlen erfüllen diese drei Kriterien: die Vorschläge von Budescu (1993) und Johnson (2000). Da die Berechnung des ersteren rechnerisch sehr aufwändig ist und die Übereinstimmung mit dem Vorschlag von Johnson groß zu sein scheint, empfehlen Chao et al. (2008) die Verwendung des letztgenannten Ansatzes. Im folgenden Absatz werden wir diesen Ansatz kurz beschreiben.

Gehen wir von einem Modell mit k Prädiktoren aus, dann beruht Johnsons Vorschlag darauf, aus diesen Merkmalen k orthogonale, also unkorrelierte, Hauptkomponenten z_m zu extrahieren und diese so zu rotieren, dass die Summe der quadrierten Abweichungen zwischen den Beobachtungswerten x_{ij} und den Faktorscores z_{im} minimiert wird. Mit den extrahierten Faktoren wird nun eine Regressionsanalyse auf die interessierende abhängige Variable gerechnet. Da die Faktoren orthogonal sind, entspricht die Summe der entsprechend standardisierten Regressionskoeffizienten $B_{z_m}^*$ dem Anteil der erklärten Varianz R^2. Nun muss noch die Bedeutung der ursprünglichen unabhängigen Variablen x_j bestimmt werden. Diese werden nach

$$B^{\dagger}_{x_j} = \sum_{m=1}^{k} \lambda_{jm} B^{*}_{z_m}$$

berechnet. λ_{jm} bezeichnet hierbei die Korrelationen bzw. Ladungen zwischen x_j und z_m.

Eine leicht verfügbare Alternative kann aus dem Beitrag von Bring (1994) abgeleitet werden. Wie er zeigt, kann die relative Bedeutung der einzelnen unabhängigen Variablen aus den t-Werten des üblicherweise verwendeten zweiseitigen Tests der Steigungskoeffizienten abgelesen werden. Da diese Prüfgröße auch als

$$t_1 = \sqrt{\frac{R^2_{1,2,3,\ldots,\,k} - R^2_{2,3,\ldots,\,k}}{(1 - R^2_{1,2,3,\ldots,\,k})/(n-k-1)}} \tag{26}$$

geschrieben werden kann, ist sie eine direkte Funktion des Zuwachses an R^2, der durch die Aufnahme der interessierenden Variable in das Modell entsteht (Bring 1994, S. 213). Folglich gibt ein Vergleich der t-Werte innerhalb desselben Modells ceteris paribus auch Auskunft über die relative Einflussstärke der unabhängigen Variablen.

Dem interessierten Nutzer bieten sich also verschiedene Alternativen zur Bestimmung der relativen Einflussstärke. Standardisierte B^*-Koeffizienten sind leicht verfügbar, aber unter Umständen problematisch. Die von Johnson vorgeschlagene Variante ist zwar weniger problematisch, aber nicht immer verfügbar. Unsere Empfehlung lautet daher, neben standardisierten unbedingt auch unstandardisierte Koeffizienten zu berichten, und eine Interpretation der relativen Einflusstärke nicht allein auf B^*-Koeffizienten zu stützen. Vielmehr sollten zusätzlich die t-Werte berücksichtigt werden.

3 Ein Beispiel

Nachdem wir die wichtigsten mathematisch-statistischen Grundlagen der linearen Regression vorgestellt haben, soll die Anwendung des Verfahrens nun an einem Beispiel diskutiert werden. Dabei werden wir untersuchen, von welchen Faktoren die Höhe des Erwerbseinkommens von abhängig Beschäftigten abhängt. Als empirische Grundlage dient uns der ALLBUS 2006. Nach der Humankapitaltheorie sollte das Einkommen vor allem von der Bildung und der Berufserfahrung abhängen. Darüber hinaus wissen wir aus vielen Arbeiten, dass Männer noch immer mehr verdienen als Frauen. Hinzu kommt, dass das Lohnniveau in den alten Bundesländern nach wie vor über demjenigen der neuen Bundesländer liegt. Aus diesen Überlegungen ergibt sich das zu schätzende Regressionsmodell

$$\widehat{\text{Einkommen}} = f(\text{Bildung, Berufserfahrung, Geschlecht, Ost/West})\,.$$

3.1 Zur Operationalisierung

Bei der Variablen „Einkommen“ handelt es sich um das persönliche monatliche Nettoeinkommen in Euro. Diese Angabe ist aus mindestens zwei Gründen für die hier

verfolgte Fragestellung nicht optimal. Erstens handelt es sich bei diesem „Einkommen“ nicht ausschließlich um Erwerbseinkommen, sondern um das gesamte persönliche Einkommen in Vollzeit tätiger abhängig Beschäftigter, also abzüglich Steuern und Sozialversicherungsbeiträgen, aber inklusive Sozialleistungen, Kapitaleinkünften, privater Transfers etc. In der untersuchten Gruppe sollte allerdings der ganz überwiegende Teil des Einkommens aus Erwerbsarbeit stammen. Zweitens beziehen sich die Aussagen der Humankapitaltheorie auf den Brutto(stunden)lohn und nicht auf den Nettolohn, der auch von anderen Faktoren, insbesondere der familiären Situation, abhängt. Aus diesem Grund werden wir in den folgenden Modellen die Anzahl der Kinder im Haushalt, den Status verheiratet versus nicht verheiratet sowie einen Interaktionseffekt zwischen dem Status verheiratet und dem Geschlecht kontrollieren (zu Interaktionseffekten siehe ausführlich Kapitel 26). Diese Faktoren kennzeichnen wesentliche, nicht direkt mit der Einkommenshöhe in Verbindung stehende Elemente der deutschen Einkommenssteuer.

Die Investitionen in Bildung werden hier operationalisiert durch eine Kombination aus dem höchsten Abschluss einer allgemeinbildenden Schule und dem höchsten beruflichen Abschluss. Die resultierende Bildungsvariable hat fünf Ausprägungen: (1) höchstens Hauptschulabschluss mit Lehre (29 %); (2) mindestens Mittlere Reife mit einer Lehre oder einem Fachschulabschluss (44 %); (3) Techniker oder Meister (7 %); (4) Fachhochschulabschluss (7 %); (5) Hochschulabschluss (13 %). Die zweite Komponente des Humankapitals, die Berufserfahrung, wird im ALLBUS – wie in den meisten Studien – nicht direkt gemessen. Für Männer wurde dieses Merkmal aus dem Alter abzüglich der in Ausbildung verbrachten Zeiten und abzüglich der ersten sechs Lebensjahre berechnet. Für Frauen wurde von dieser Zahl noch einmal jeweils drei Jahre für jedes Kind abgezogen. Die Berufserfahrung wird hier in Dekaden gemessen und um ihren Mittelwert zentriert. Die Merkmale „Geschlecht“ und „alte vs. neue Bundesländer“ werden als Dummy-Variablen in die Analyse eingeführt. Sie sind so kodiert, dass die ausgewiesenen Effekte für Männer bzw. Personen in Westdeutschland gelten.

3.2 Ergebnisse

Modell 1 in Tabelle 1 enthält die bisher vorgestellten Merkmale. Die beiden Indikatoren des Humankapitals zeigen die erwarteten Ergebnisse. Je höher der erreichte Ausbildungsabschluss und je umfangreicher die Berufserfahrung, umso höher ist das erwartbare Einkommen. Vollzeit Erwerbstätige, die höchstens einen Hauptschulabschluss mit Lehre aufweisen, verdienen 339 € weniger als Erwerbstätige mit Mittlerer Reife und Lehre bzw. Fachschulausbildung, 415 € weniger als Techniker und Meister, 889 € weniger als Fachschulabsolventen und sogar 1362 € weniger als Erwerbstätige mit Hochschulabschluss. Unabhängig vom Qualifikationsniveau führt die Berufserfahrung in zehn Jahren zu einer durchschnittlich zu erwartenden Einkommenserhöhung von 139 €. Allerdings postuliert die Humankapitaltheorie, dass das Einkommen nicht linear mit der Berufserfahrung steigt. Vielmehr wird ein abnehmender Grenzertrag zunehmender Erfahrung erwartet. Diese Vorstellung kann in unsere Analyse einfließen, indem wir die Berufserfahrung auch quadriert in die Analyse aufnehmen (siehe dazu auch Kapitel 26). Die Analyse bleibt dennoch eine lineare Regressionsanalyse, weil

sie nach wie vor linear in ihren Parametern ist. Der entsprechende Ausschnitt aus der Regressionsgleichung lautet folglich

$$\widehat{\text{Einkommen}} = \cdots \hat{\beta}_3 \text{ Bildung} + \hat{\beta}_4 \text{ Erfahrung} + \hat{\beta}_5 \text{ (Erfahrung)}^2 \cdots .$$

Wie eine entsprechende Analyse zeigt (nicht abgedruckt), hat der quadrierte Term zwar das erwartete negative Vorzeichen, d. h. die erfahrungsbedingten Einkommenszuwächse werden mit steigender Erfahrung kleiner. Allerdings ist dieser Effekt mit lediglich 8 € in der ersten Dekade, 16 € in der zweiten Dekade, 72 € in der dritten Dekade schwach und auch unter statistischen Gesichtspunkten bedeutungslos. Daher werden wir diesen Term nicht weiter berücksichtigen.

Bei der verwendeten Kodierung der Merkmale gibt die Regressionskonstante von 635 € den monatlichen zu erwartenden Nettoverdienst einer in Ostdeutschland abhängig beschäftigten, nicht verheirateten Frau wieder, die keine Kinder im Haushalt hat, höchstens über einen Hauptschulabschluss mit Lehre verfügt und eine durchschnittliche Berufserfahrung[11] von 20,8 Jahren hat. Für ihre Kollegin im Westen wird aufgrund des Modells ein um 557 € höherer Durchschnittsverdienst, also fast das Doppelte, erwartet. Vergleicht man den Verdienst von Männern und Frauen zeigen sich hier ebenfalls beträchtliche Differenzen: Unverheiratete Männer verdienen durchschnittlich 199 € mehr als entsprechende Frauen. Bei Verheirateten beträgt die Differenz sogar 529 €. Um zu verstehen, wie sich diese Angaben berechnen, sei kurz auf den entsprechenden Ausschnitt aus der Regressionsgleichung eingegangen (alle Angaben aus Modell 1 in Tabelle 1):

$$\widehat{\text{Einkommen}} = \cdots 199 \cdot \text{Mann} - 90 \cdot \text{verheiratet} + 330 \cdot \text{Mann} \cdot \text{verheiratet} \cdots$$

Für das Geschlecht und den Familienstand berücksichtigen wir je einen Haupteffekt und zusätzlich den Interaktionseffekt der beiden Merkmale. Unverheiratete Frauen stellen unseren Bezugspunkt, unsere Referenzkategorie, dar. Ein unverheirateter Mann verdient 199 € mehr als eine unverheiratete Frau. Eine verheiratete Frau verdient 90 € weniger als eine unverheiratete Frau. Ein verheirateter Mann verdient durchschnittlich 330 € mehr als ein unverheirateter Mann und 529 € (=199+330) mehr als eine unverheiratete Frau. Im Vergleich zu einer verheirateten Frau verdient ein verheirateter Mann sogar 619 € (529+90) mehr.

All diese Angaben sind bedingte Erwartungen für das durchschnittliche Einkommen der genannten Personengruppen bei ansonsten gleichen Merkmalen, hier also gleicher Bildung und gleicher Berufserfahrung. Weil die Bildung und die Berufserfahrung im Modell bereits kontrolliert sind, sind die Unterschiede zwischen West- und Ostdeutschland, aber auch die Geschlechterdifferenz besonders eklatant. In Bezug auf die Einkommensunterschiede zwischen Ost und West könnte man allerdings argumentieren, dass die in der DDR erworbenen Ausbildungsabschlüsse sowie die dort gemachte Berufserfahrung im wiedervereinigten Deutschlands nicht ebenso produktiv sind wie das entsprechende westdeutsche Humankapital.[12]

[11] Wie oben erläutert, ist das Merkmal Berufserfahrung in den hier präsentierten Analysen um seinen Mittelwert zentriert.

[12] Auf der Basis von Analysen, die sich nur auf Personen beschränken, die 1990 höchstens 18 Jahre alt waren, ihre Bildung und Berufserfahrung also nach der Wiedervereinigung erwor-

Tab. 1: Regressionsanalysen des Einkommens

	Modell 1				Modell 2			
	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t
Konstante	635	80		7,91	740	114		6,51
Westen	557	53	0,27	10,50	513	52	0,25	9,82
Männlich	199	66	0,11	3,02	215	64	0,12	3,36
verheiratet	−90	73	−0,05	−1,24	−84	71	−0,05	−1,18
Mann × verheiratet	330	90	0,19	3,66	349	87	0,21	4,00
Kinder	65	24	0,08	2,70	73	23	0,09	3,13
Bildung (Ref. HS, Lehre)								
MR, Lehre	339	52	0,20	6,58	210	53	0,13	3,98
Techn./Meister	415	90	0,12	4,60	228	91	0,07	2,51
FH	889	90	0,26	9,83	569	97	0,17	5,86
Uni	1362	70	0,54	19,53	895	91	0,36	9,82
Berufserfahrung	139	22	0,18	6,43	125	21	0,16	5,95
Berufsprestige					71	10	0,25	7,48
Deutsch					72	78	0,02	0,93
R^2	0,46				0,49			
R^2_{korr}	0,45				0,49			

Datenbasis: ALLBUS 2006; gewichtet mit Ost-West Transformationsgewicht (n=907).
Nur ganztags Erwerbstätige mit abhängiger Beschäftigung.

Ein Merkmal, welches in der Einkommensgleichung von Modell 1 noch nicht berücksichtigt wird, aber nach soziologischen Theorien eine Rolle spielen sollte, ist der ausgeübte Beruf. Nach den soziologischen Theorien des Statuserwerbs hat die Bildung zunächst einen Einfluss auf den Status des ausgeübten Berufs und dieser wiederum beeinflusst das Einkommen (vgl. Blau & Duncan 1967). Daher wurde in Modell 2 zusätzlich das Berufsprestige aufgenommen.[13] Dieses Merkmal hat den erwarteten starken Effekt auf das Einkommen. Zwischen Berufen, die 10 Punkte auf der Prestigeskala auseinander liegen, wird ein durchschnittlicher, bedingter Einkommensunterschied von 71 € erwartet. Bei einer Spannweite des Berufsprestiges von 166,8 (=186,8-20,0) Punkten, ergibt sich eine bedingte Einkommensdifferenz zwischen Personen mit dem höchsten und Personen mit dem niedrigsten Berufsprestige von 1193 € ($= 71 \cdot {}^{(186{,}8-20{,}0)}/_{10}$). Wie wirkt sich die Aufnahme des Berufsprestiges auf den Einfluss der anderen Merkmale aus? Wie ein Vergleich des Bildungseffekts in Modell 2

ben haben, zeigt sich jedoch nach wie vor ein großer Einkommensunterschied zwischen den beiden Landesteilen. Da die jungen Ostdeutschen bereits das Bildungssystem des wiedervereinigten Deutschland durchlaufen haben, kann das angeführte Humankapitalargument nicht zur Begründung von Einkommensdifferenzen bemüht werden.

[13] Das Berufsprestige wurde hier nach der Magnitude-Prestigeskala von Wegener (1988) gemessen, eine im ALLBUS bereits vorhandene Variable. Die ursprüngliche Skala, die von 20 bis 186,8 Punkten reicht, wurde für die in Tabelle 1 präsentierte Analyse zentriert und durch 10 dividiert.

mit dem in Modell 1 zeigt, verringern sich die Einkommensunterschiede zwischen den Bildungsgruppen, wenn das Berufsprestige kontrolliert wird. Das bedeutet, dass es sich bei einem Teil des in Modell 1 ausgewiesenen Bildungseffekts auf das Einkommen um einen indirekten Effekt handelt. Die Einflussstärke der anderen unabhängigen Variablen bleibt dagegen im Wesentlichen unverändert.

Eine weitere Variable, die mit dem Einkommen in Verbindung stehen könnte, ist die Nationalität. Häufig wird die Vermutung geäußert, dass Ausländer auf dem Arbeitsmarkt diskriminiert werden und weniger verdienen als Deutsche. Die in Modell 2 von Tabelle 1 wiedergegebene Analyse stützt diese Hypothese nicht. Die Nettoeinkommen von Deutschen und Ausländern unterscheiden sich nicht signifikant. Zwar liegt der Erwartungswert für das Nettoeinkommen der Deutschen bei gleicher familialer Situation, gleicher Bildung, gleicher Berufserfahrung etc. um 72 € höher bei Ausländern. Mit einem Standardfehler von 78 € ist dieser Effekt jedoch sehr weit von jeglicher statistischer Signifikanz entfernt. Dies deckt sich mit früheren Ergebnissen, die ebenfalls keinen Einkommensnachteil (*ethnic penalty*) von Ausländern ermitteln konnten (Diekmann et al. 1993).

Welches der untersuchten Merkmale hat den stärksten Einfluss auf das Einkommen? Nach den standardisierten Koeffizienten zu urteilen, ist es das Vorhandensein eines Universitätsabschlusses, gefolgt vom Wohnen in Westdeutschland und dem Berufsprestige, die mit einem Koeffizient von jeweils 0,25 gleichauf sind. Dieser Vergleich ist jedoch irreführend und zwar unabhängig von den in Abschnitt 2.7 beschriebenen Problemen der standardisierten Koeffizienten. Die Effektstärke einer kategorialen Variablen, die durch mehrere Dummy-Variablen repräsentiert wird, kann nicht an den einzelnen standardisierten Effekten abgelesen werden. Auch die t-Werte geben keine Auskunft über die statistische Bedeutung des (mehrstufig kategorialen) Merkmals als Ganzes. Um festzustellen, ob ein solches Merkmal einen statistisch signifikanten Einfluss auf die untersuchte abhängige Variable hat und wie stark dieser Einfluss ist, muss ein Modell, in dem die entsprechenden Dummies enthalten sind, mit einem Modell verglichen werden, in dem die Dummies nicht enthalten sind. Entfernt man die vier Bildungs-Dummies aus Modell 2 (Tabelle 1), dann sinkt die erklärte Varianz um über sieben Prozentpunkte; eine sowohl unter substanziellen als auch unter statistischen Gesichtspunkten signifikante Verringerung. Zum Vergleich: Wird die Region aus dem Modell entfernt, sinkt die erklärte Varianz um fünf Punkte, beim Berufsprestige um drei Punkte und bei der Berufserfahrung um zwei Punkte. Einen Vergleich mit dem Geschlecht können wir hier leider nicht vornehmen, da aus den obengenannten Gründen auch ein Interaktionsterm zwischen Geschlecht und Familienstand im Modell enthalten ist. Werden alle drei steuerlich relevanten Merkmale ausgeschlossen – Geschlecht, Familienstand, Kinder –, dann verringert sich das R^2 um acht Punkte. Unter den analysierten Merkmalen sind demnach die Bildung und die Region des Wohnorts die bedeutsamsten Determinanten des Einkommens in Deutschland.

Die Einkommensvariable ist in der Regel rechtsschief verteilt, da verhältnismäßig viele Personen wenig, wenige Personen hingegen sehr viel verdienen. Das Einkommen ist also nicht normalverteilt. Dies hat meist zur Folge, dass auch die Residuen nicht normalverteilt sind und damit eine Anwendungsvoraussetzung der Kleinst-Quadrat-Methode nicht gegeben ist. Diese Annahmeverletzung kann zu verzerrten Standardfehlern und

Tab. 2: Regressionsanalysen des Einkommens, metrische versus logarithmierte Einkommensvariable

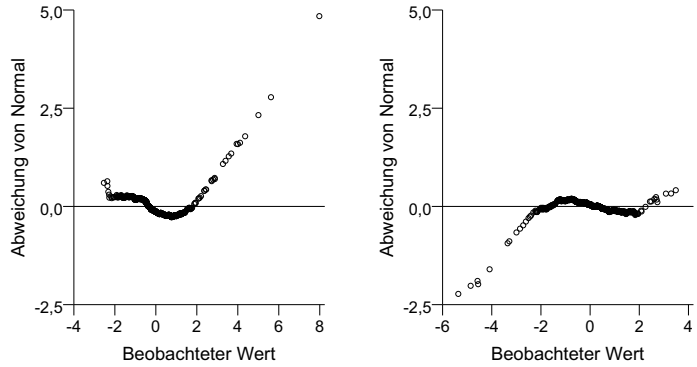
	Modell 1: Einkommen in Euro				Modell 2: Logarithmus des Einkommens			
	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t	$\hat{\beta}$	$s_{\hat{\beta}}$	B^*	t
Konstante	814	81		10,01	6,81	0,044		155,25
Westen	507	52	0,24	9,78	0,32	0,028	0,29	11,52
Männlich	214	64	0,12	3,35	0,11	0,035	0,12	3,29
verheiratet	−86	71	−0,05	−1,22	−0,07	0,038	−0,08	−1,88
Mann × verheiratet	347	87	0,20	3,98	0,21	0,047	0,23	4,41
Kinder	72	23	0,08	3,07	0,04	0,013	0,09	3,12
Bildung (Ref. HS, Lehre)								
MR, Lehre	212	53	0,13	4,03	0,16	0,028	0,17	5,51
Meister	236	91	0,07	2,60	0,20	0,049	0,11	4,17
FH	573	97	0,17	5,90	0,33	0,052	0,18	6,39
Uni	897	91	0,36	9,85	0,45	0,049	0,34	9,26
Berufserfahrung	126	21	0,17	6,02	0,08	0,011	0,20	7,28
Berufsprestige	72	9	0,25	7,62	0,03	0,005	0,22	6,49
R^2	0,49				0,49			
R^2_{korr}	0,49				0,48			

Datenbasis: ALLBUS 2006; gewichtet mit Ost-West Transformationsgewicht (Fallzahl 907). Nur ganztags Erwerbstätige mit abhängiger Beschäftigung.

damit zu falschen Schlüssen aus Signifikanztests führen. Eine Lösung dieser Problematik kann in der Transformation der abhängigen Variablen bestehen. Im Falle von rechtsschiefen Merkmalen, wie dem Einkommen, führt das Logarithmieren oftmals zu einer angemesseneren Verteilung. Aus diesem Grund verwendet man bei Einkommensanalysen standardmäßig das logarithmierte Einkommen. Ein solches Modell wurde daher auch hier gerechnet und soll nun mit den bisher erzielten Ergebnissen verglichen werden (vgl. Tabelle 2). Beide Modelle beinhalten dieselben unabhängigen Variablen, und zwar solche, die nach den bisher durchgeführten Analysen einen statistisch bedeutsamen Beitrag zur Erklärung des Einkommens liefern (vgl. Tabelle 1 auf Seite 631). In Modell 1 wurde als abhängige Variable wieder das Nettoeinkommen in Euro verwendet. In Modell 2 dient der natürliche Logarithmus des Nettoeinkommens als abhängige Variable. Um die ausgewiesenen Effekte interpretieren zu können, muss die Exponentialfunktion angewandt werden. Damit ergibt sich

$$\widehat{\text{Einkommen}} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \cdots e^{\beta_k x_k}$$

als zu schätzende Regressionsgleichung. Eine Erhöhung von x_1 um eine Einheit führt in diesem Modell zu einer Veränderung des Einkommens um den Faktor β_1. Da für Exponenten $c \leq 0{,}2$ gilt, dass $e^c \approx 1+c$, werden Regressionskoeffizienten mit einer logarithmierten abhängigen Variablen häufig als prozentuale Veränderungen interpretiert.



(a) Einkommen in Euro

(b) logarithmiertes Einkommen

Abb. 3: Trendbereinigte Q-Q-Plots der standardisierten Residuen

Dies ist gerade in ökonomischen Analysen für das Einkommen sehr beliebt. Nehmen wir beispielweise den Effekt der Berufserfahrung in Höhe von 0,08 (vgl. Tabelle 2, Modell 2). Da $e^{0,08} = 1,08$ ist, kann dieser Effekt dahingehend interpretiert werden, dass Erwerbstätige mit zehn Jahren mehr Berufserfahrung als andere Erwerbstätige ein um durchschnittlich 8 % höheres Einkommen aufweisen. Ist der Regressionskoeffizient deutlich größer als 0,2, dann kann das Ergebnis von e^{β} nicht direkt an β abgelesen werden. Nehmen wir beispielsweise die Effekte für einen Fachhochschul- (0,33) oder Universitätsabschluss (0,45). Für den Fachhochschulabschluss ergibt sich $e^{0,33} = 1,45$ – und eben nicht nicht 1,33. Hinsichtlich des Universitätsabschlusses lautet das Ergebnis $e^{0,45} = 1,57$ (also deutlich höher als 1,45). Allgemein gilt: Je größer c ist, desto stärker weicht die tatsächliche prozentuale Veränderung von diesem Wert ab.

Im Großen und Ganzen führen beide Analysen zum selben Ergebnis. Beide belegen die großen Einkommensunterschiede zwischen West- und Ostdeutschland, die Einkommensdifferenz zwischen Männern und Frauen, den starken Bildungseffekt und die Effekte von Berufserfahrung und Berufsprestige. Zudem erklären beide Modelle denselben Anteil an Einkommensvarianz. Betrachtet man hingegen die Standardfehler, kann festgestellt werden, dass das Modell mit logarithmierter abhängiger Variable relativ zu den Effektstärken meist kleiner sind.

Zusammengenommen ergibt sich in dieser Beispielanalyse kein großer Vorteil aus der Verwendung des logarithmierten Einkommens. Dies kann jedoch in anderen (vor allem kleineren) Stichproben und insbesondere bei einer anderen Operationalisierung des Einkommens anders sein. Im Gegensatz zu den meisten ökonomischen Einkommensanalysen untersuchen wir hier nicht das Brutto-, sondern das Nettoeinkommen. Letzteres ist aufgrund der abgezogenen Einkommenssteuer und dem hinzu gezählten Transfereinkommen deutlich weniger rechtsschief verteilt als das Bruttoeinkommen. Eine Transformation ist deshalb in diesem Fall weniger „nötig“. Eine Analyse der Feh-

lerterme der beiden in Tabelle 2 dargestellten Modelle macht dies deutlich. Abbildung 3 bietet trendbereinigte Q-Q-Plots der standardisierten Residuen aus beiden Modellen. Diese Plots zeigen, wie stark die Residuen von einer Normalverteilung abweichen. Bei einer perfekt normalverteilten Variablen lägen alle Punkte auf der eingezeichneten horizontalen Linie. Abweichungen von dieser Linie nach unten oder oben zeigen entsprechende Abweichungen von der Normalverteilung an (vgl. ausführlich dazu die Kapitel 5 und 25 in diesem Handbuch). Wie sich aus Abbildung 3 ersehen lässt, weichen die beiden Verteilungen in unterschiedlicher Form von der Normalverteilung ab. Beim nicht transformierten Einkommen treten die Abweichungen vor allem im Bereich hoher Einkommen auf; beim logarithmierten Einkommen finden sich die Abweichungen dagegen am unteren Ende der Einkommensverteilung. Das Logarithmieren behebt demnach das Problem bei den hohen Einkommen (die Rechtsschiefe), führt jedoch zu einer größeren Abweichung bei den geringen Einkommen. Insgesamt sind die Abweichungen in beiden Fällen jedoch verhältnismäßig gering.

4 Häufige Fehler

Wie bei allen statistischen Verfahren kann eine sachlich angemessene Interpretation von Ergebnissen der linearen Regression nur erfolgen, wenn die mathematisch-statistischen Grundlagen und Annahmen sowie die grundlegende Funktionsweise des Verfahrens in seinen Grundzügen verstanden wurden. Eine dieser Grundannahmen ist, dass die untersuchten Prädiktoren in einem linearen Zusammenhang mit der abhängigen Variablen stehen. Diese Annahme sollte in jedem einzelnen Fall überprüft werden. Dies kann auf mindestens zweierlei Weise erfolgen. Einerseits kann die unabhängige Variable in mehrere Gruppen unterteilt werden, die dann als Dummy-Variablen in das Regressionsmodell aufgenommen werden können. Anhand der Regressionskoeffizienten lässt sich leicht ablesen, ob die Linearitätsannahme gerechtfertigt ist. Andererseits kann die fragliche unabhängige Variable in einem Streudiagramm gegen die abhängige Variable geplottet und die Regressionsgerade mit einer nichtparametrischen lokal gewichteten Regressionskurve (LOWESS) verglichen werden (siehe Kapitel 25 in diesem Handbuch).

Gelangt man zu dem Schluss, dass Nichtlinearität vorliegt, kann dem oftmals durch die Berücksichtigung von Polynomen Rechnung getragen werden (vgl. Kapitel 26 in diesem Handbuch). Die Verwendung von Polynomen, am häufigsten quadrierte Terme, führt jedoch potenziell zu einer weiteren Schwierigkeit, dem Problem hoher Multikollinearität. Mit diesem Begriff wird die multiple Korrelation zwischen den Prädiktoren bezeichnet. Ist diese zu hoch, führt dies zur Verzerrung der Standardfehler der Regressionskoeffizienten. Gerade in kleineren Stichproben ist es dann schwer, signifikante Ergebnisse zu identifizieren. Die Höhe der Multikollinearität lässt sich an der sog. Toleranz bzw. dem Varianzinflationsfaktor ablesen (vgl. Gleichungen (23) und (24) in Kapitel 25, S. 655). Häufig kann das Problem verringert werden, wenn die Polynome auf Basis von um den Mittelwert zentrierten Merkmalen berechnet werden. Dies gilt im Übrigen auch für (andere) Interaktionseffekte.

Nicht spezifisch für die Regressionsanalyse, aber deshalb nicht weniger wichtig, ist die sorgfältige Berechnung und Prüfung aller an der Analyse beteiligten Variablen. Sind die Verteilungen plausibel? Ist der Anteil fehlender Werte nachvollziehbar? Insbesondere der Umgang mit letzteren sollte bei der Regressionsanalyse gut bedacht sein. Die Statistikprogramme, mit denen entsprechende Analysen durchgeführt werden, haben alle eine bestimmte Voreinstellung, wie sie mit Fällen umgehen, die fehlende Werte aufweisen. Diese Voreinstellung ist meistens der Ausschluss aller Fälle mit mindestens einem fehlenden Wert (*listwise deletion*). Dies kann dazu führen, dass sich die Fallzahl deutlich reduziert. Daher sollte immer überprüft werden, auf welcher Basis die eigentliche Analyse durchgeführt wird. Hat sich die Stichprobe aufgrund fehlender Werte zu sehr verringert, muss über alternative Wege im Umgang mit fehlenden Werten nachgedacht werden (vgl. Kapitel 6 in diesem Handbuch).

Ein weiteres generelles Problem besteht in der Verwechslung von statistischer und inhaltlicher Bedeutsamkeit. Ist ein bestimmter Regressionskoeffizient statistisch „signifikant“, sagt das noch nichts über die inhaltliche Bedeutung dieses Effekts aus. Auf Basis einer sehr großen Stichprobe legen bereits sehr kleine Effekte den Schluss nahe, dass der Effekt in der Grundgesamtheit von null verschieden ist. Damit wird der Effekt selbst jedoch nicht größer. Umgekehrt kann ein Koeffizient aus einer kleinen Stichprobe das Kriterium der statistischen Signifikanz zwar knapp verfehlen, aufgrund seiner Größe dennoch als ein bedeutsamer Effekt interpretiert werden. Statistische Signifikanz und inhaltliche Bedeutung sind demnach zwei verschiedene Dinge, die nicht miteinander verwechselt werden sollten.

Ein letzter hier zu nennender Komplex betrifft die Gefahr einer grundlegenden Fehlinterpretation der Ergebnisse von Querschnittsregressionen. Erstens sollte immer bedacht werden, dass sich die Koeffizienten letztlich immer auf Gruppenunterschiede oder, genauer, Unterschiede in bedingten Erwartungswerten beziehen. Für zwei Personen, die sich in Bezug auf die unabhängige Variable um eine Einheit unterscheiden, beträgt die Differenz in den bedingten Erwartungswerten der abhängigen Variablen β Einheiten. Die Formulierung „bedingte Erwartungswerte“ bezieht sich dabei darauf, dass die Regressionskoeffizienten unter Konstanthaltung der anderen berücksichtigten Merkmale verglichen werden. Die Aussage über die Differenz der Erwartungswerte gilt also *ceteris paribus* – unter sonst gleichen Umständen. Da diese Interpretation sprachlich recht umständlich ist, wird häufig – auch in diesem Beitrag – eine elegantere, aber unpräzise Formulierung gewählt: „β gibt an, um wie viele Einheiten sich die abhängige Variable verändert, wenn die unabhängige Variable um eine Einheit steigt.“ Diese Aussage verweist sprachlich auf eine Prognose, die jedoch auf Basis von Querschnittsregressionen nur unter bestimmten Voraussetzungen möglich ist.[14] Ein zweites Problem betrifft die kausale Interpretation von Regressionsergebnissen. Ob dies möglich ist, hängt nicht vom Analyseverfahren, hier also der Regression, sondern wesentlich davon ab, ob die entsprechenden Voraussetzungen für die Beobachtung eines kausalen Effektes gegeben sind. Zu diesen Voraussetzungen gehört insbesondere, dass die vermeintliche Ursache der Wirkung vorausgeht und dass alle relevanten Störgrößen

[14] Vorhersagen sind zwar auf Basis von Regressionsanalysen prinzipiell möglich, setzen aber einer Erweiterung des Verfahrens voraus (vgl. Cohen et al. 2003, S. 95 ff.).

kontrolliert werden. Diese Voraussetzungen werden am ehesten unter experimentellen Bedingungen erfüllt (vgl. ausführlich die Kapitel 2, 35 und 36 in diesem Handbuch).

5 Literaturhinweise

Das Verfahren der linearen Regressionsanalyse wird in nahezu jedem Lehrbuch zur Statistik behandelt. Darüber hinaus gibt es unzählige monographische Darstellungen dieses Verfahrens. Einen guten Einstieg bieten die Bücher von Urban & Mayerl (2006) sowie Gelman & Hill (2007). Eine leicht verständliche Einführung in die Voraussetzungen der linearen Regression und ihre Bedeutung liefert Berry (1993). Wer einen kürzeren Überblicksartikel zum Verfahren sucht, dem sei der Beitrag von Stolzenberg (2004) empfohlen. Eine didaktisch hervorragende und mathematisch präzise Darstellung bietet Wooldridge (2009), vertiefende Ausführungen findet man bei Wooldridge (2002). Ein Aspekt, auf den wir nicht eingehen konnten, betrifft die spezifischen Probleme der Regressionsanalyse bei kleinen Stichproben. Dieser Thematik widmet sich Jann (2009).

Literaturverzeichnis

Bacher, J. (2009). Analyse komplexer Stichproben. In M. Weichbold, J. Bacher, & C. Wolf (Hg.), *Umfrageforschung. Herausforderungen und Grenzen*, Band 9 (S. 253–274). Wiesbaden: VS Verlag für Sozialwissenschaften.

Berry, W. D. (1993). *Understanding Regression Assumptions*, Band 07-092 von *Quantitative Applications in the Social Sciences*. Newbury Park: Sage.

Best, H. (2009). Organic Farming as a Rational Choice. Empirical Investigations in Environmental Decision Making. *Rationality and Society*, 21, 197–224.

Blau, P. M. & Duncan, O. D. (1967). *The American Occupational Structure*. New York: Wiley.

Bring, J. (1994). How to Standardize Regression Coefficients. *The American Statistician*, 48, 209–213.

Budescu, D. V. (1993). Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression. *Psychological Bulletin*, 114, 542–551.

Chao, Y.-C. E., Zhao, Y., Kupper, L. L., & Nylander-French, L. A. (2008). Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. *Journal of Occupational and Environmental Hygiene*, 5, 519–529.

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah: Lawrence Erlbaum, 3. Auflage.

Diekmann, A., Engelhardt, H., & Hartmann, P. (1993). Einkommensungleichheit in der Bundesrepublik Deutschland: Diskriminierung von Frauen und Ausländern? *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 3/93, 386–398.

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistican*, 61, 139–147.

Jann, B. (2009). Diagnostik von Regressionsschätzungen bei kleinen Stichproben (mit einem Exkurs zu logistischer Regression). In P. Kriwy & C. Gross (Hg.), *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen* (S. 93–126). Wiesbaden: VS Verlag für Sozialwissenschaften.

Johnson, J. W. (2000). A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. *Multivariate Behavioral Research*, 35, 1–19.

Lee, E. S. & Forthofer, R. N. (2006). *Analyzing Complex Survey Data*, Band 07-071 von *Quantitative Applications in the Social Sciences*. Thousand Oaks: Sage, 2. Auflage.

Stolzenberg, R. M. (2004). Multiple Regression Analysis. In M. Hardy & A. Bryman (Hg.), *Handbook of data analysis* (S. 165–208). London: Sage Publications.

Urban, D. & Mayerl, J. (2006). *Regressionsanalyse: Theorie, Technik und Anwendung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 2. Auflage.

Wegener, B. (1988). Die Magnitude-Prestigeskala (MPS) - Theorie, Konstruktion und die Prestigescores für berufliche Tätigkeiten. In B. Wegener (Hg.), *Kritik des Prestige* (S. 221–244). Opladen: Westdeutscher Verlag.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wooldridge, J. M. (2009). *Introductory Econometrics. A Modern Approach*. o.O.: South-Western, 4. Auflage.