Grundlagen der Regressionsanalyse

2

Lernziele
Der Studierende soll:

- die Methode der kleinsten Quadrate als Schätzverfahren zu Ermittlung des Zusammenhangs zwischen einer abhängigen und mehreren unabhängigen Variablen erläutern können,
- wissen, wie die Koeffizientenschätzungen im Rahmen einer einfachen und einer multiplen linearen Regressionsanalyse zu interpretieren sind,
- den Output einer multiplen Regression der Statistikprogrammpakete SPSS sowie Stata und damit prinzipiell auch anderer Statistikprogramme verstehen,
- den Unterschied zwischen einer beschreibenden (deskriptiven) und einer induktiven Verwendung und Interpretation der Regression darlegen können,
- erläutern können, dass die Schätzung der abhängigen Variablen $\hat{Y}$ die Schätzung eines bedingten Mittelwerts von Y darstellt,
- die Wichtigkeit der Aussage der Nullhypothese eines statistischen Tests verstehen,
- in der Lage sein, das Vorgehen bei der Überprüfung einer Nullhypothese zu erläutern,
- zwischen statistischer Signifikanz und inhaltlicher Relevanz eines Koeffizienten differenzieren,
- überblicken, wie Beta-Koeffizienten zu interpretieren sind.

► **Wichtige Grundbegriffe** Schätzung, Spezifikation, OLS, KQ-Methode, Konstante, t-Wert, F-Wert, R^2, korrigiertes R^2, Korrelationskoeffizient, linearer Zusammenhang, Beta-Werte, Bestimmtheitsmaß, Residuum, Sample, Nullhypothese, empirisches Signifikanzniveau, Irrtumswahrscheinlichkeit

M.-W. Stoetzer, *Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 1*, DOI 10.1007/978-3-662-53824-1_2

2.1 Überblick

Der Abschn. 2.2 erläutert das Grundprinzip der Regressionsanalyse an dem bereits bekannten PKW-Beispiel. Anschließend beschäftigt sich Abschn. 2.3 mit den grundlegenden Fragen, die bei der Beurteilung der Aussagekraft von Regressionsergebnissen zu überprüfen sind. Abschn. 2.4 fasst die wichtigsten Aussagen zusammen und Abschn. 2.5 zeigt anhand der Programmpakete SPSS und Stata wie eine solche einfache Regression durchgeführt wird.

2.2 Einfache und multiple Regression

Die Regressionsanalyse untersucht den Zusammenhang zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen hinsichtlich der Frage, ob überhaupt ein Zusammenhang besteht und wie stark er gegebenenfalls ist. Bei lediglich einer unabhängigen Variablen handelt es sich um eine **einfache Regression**, bei mehreren unabhängigen Variablen um eine **multiple Regression**.[1]

Im PKW-Beispiel des Kap. 1 geht es um die Nachfrage nach einem Produkt eines PKW-Produzenten (Menge verkaufter PKW pro Periode). Das bereits aufgeworfene Problem ist, inwieweit die Absatzmenge von der Zahl der Kontakte abhängt. Unter inhaltlichen Gesichtspunkten kann angenommen werden, dass die Zahl der Kontakte die Absatzzahlen positiv beeinflusst. Die Hypothese zur Ursache-Wirkungsbeziehung ist, dass von der Variable X_1 (*Kontakte*) eine positive Wirkung auf die verkauften PKW, die Variable Y (*Menge*), ausgeht. Je häufiger ein Produktmanager die Händler in der Region persönlich aufsucht, desto größer ist der Absatz.

Ein naheliegender Einwand ist, dass dieser Zusammenhang doch selbstverständlich ist, also überhaupt nicht untersucht werden muss. Dazu sind vier Aspekte ins Feld zu führen. Erstens ist die beschriebene Wirkung nicht zweifelsfrei. Aufgrund von schlecht ausgebildetem bzw. motiviertem Vertriebspersonal oder weil sich die PKW von alleine verkaufen, könnte auch keinerlei Einfluss vorhanden sein. Zweitens ist die genaue Höhe (das Ausmaß) der Wirkung eines Vertreterkontaktes betriebswirtschaftlich relevant. Drittens ist es möglich, dass die Auswirkungen im vorliegenden Datensatz zufälliger Natur sind und daher lediglich in genau dieser Stichprobe vorliegen. Viertens ist denkbar, dass andere Einflussfaktoren für die unterschiedlichen Verkaufsmengen verantwortlich sind. Alle vier Gesichtspunkte werden in den anschließenden Ausführungen behandelt.

Der Zusammenhang der Variablen *Menge* und *Kontakte* kann in der einfachsten Form deskriptiv graphisch untersucht werden. In das Koordinatensystem der Abb. 2.1 werden

[1] Der Begriff multivariate Regression (bzw. multivariate Analyse) wird im Folgenden (so wie in der Literatur üblich) nur für Untersuchungen verwendet, bei denen gleichzeitig *mehrere abhängige* Variablen existieren. In manchen Lehrbüchern wird davon abweichend das, was hier multiple Regression genannt wird, als multivariate Regression bezeichnet (so bspw. in den Lehrbüchern von Backhaus et al. (2011, 2013) und Studenmund (2014)).

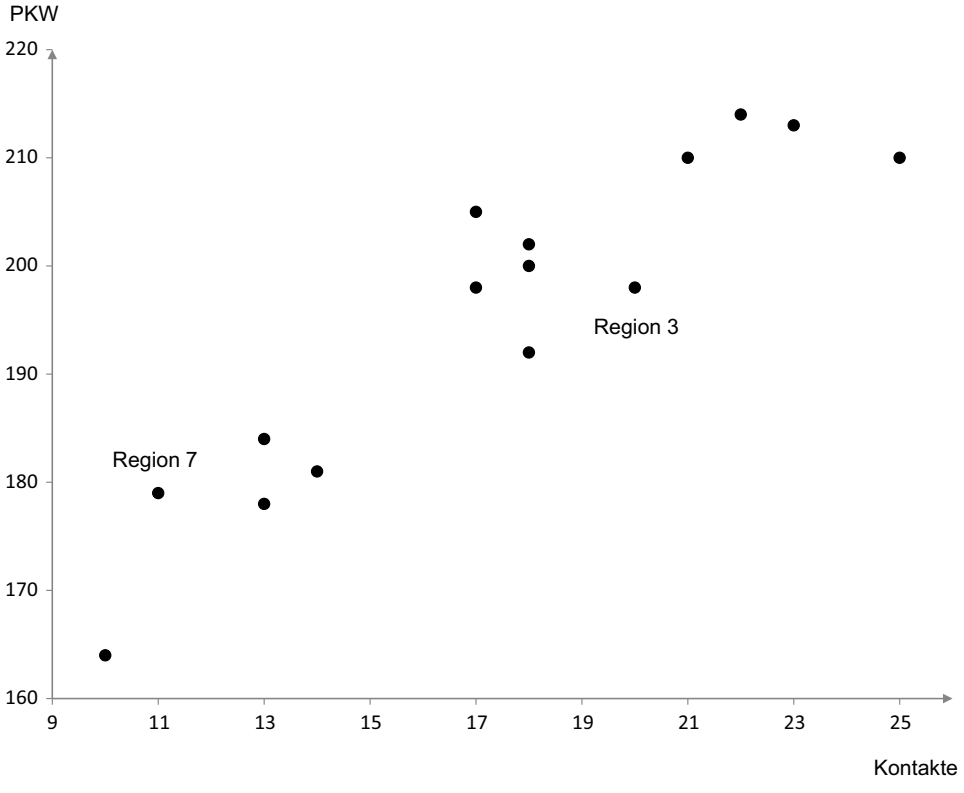

Abb. 2.1 Der Einfluss der Kontakte

die Werte von Y und X_1 aus Tab. 1.2 in Form eines Streudiagramms eingetragen. Beispielhaft sind die Beobachtungen für die Region 3 (198 verkaufte PKW bei 20 Kontakten) und die Region 7 (179 PKW und 11 Kontakte) identifiziert. Es wird deutlich, dass die verkauften Stückzahlen mit der Zahl der Kontakte steigen. Dieser Zusammenhang existiert allerdings nur tendenziell und die genaue Stärke ist in der grafischen Betrachtung schwer abzuschätzen. Die exakte Höhe des Einflusses ist aber wichtig, um entscheiden zu können, ob die Verringerung oder Erhöhung der Zahl der persönlichen Vertreterkontakte und damit indirekt auch die Entlassung oder Einstellung von Produktmanagern unter Kostengesichtspunkten für das Unternehmen notwendig bzw. möglich ist.

Der Zusammenhang von Kontakten und Absatzmenge kann präziser mit der Methode der Regressionsanalyse ermittelt werden. Der **allgemeine Zusammenhang** zwischen Y (verkaufte PKW) als abhängiger und X_1 (Kontakte) als unabhängiger Variable lautet:

$$Y = f(X_1). \tag{2.1}$$

Gl. 2.1 sagt aber noch nichts über die exakte funktionale Beziehung zwischen Y und X_1. Diese wird genauer formuliert als **einfache lineare Regression** der Art:

$$Y = B_0 + B_1 X_1 + E. \tag{2.2}$$

Diese Formulierung eines konkreten Zusammenhangs von Y und X_1 wird als **Spezifikation** bezeichnet. Gl. 2.2 stellt den unbekannten wahren Zusammenhang von X_1 und Y dar.

Y = abhängige Variable: *Menge*
B_0 = konstantes (absolutes) Glied
B_1 = Regressionskoeffizient der Variable X_1
X_1 = unabhängige Variable: *Kontakte*
E = Fehler (error term)

In Bezug auf Y und X_1 sind Daten vorhanden (siehe Tab. 1.2). Die Koeffizienten (= Parameter) B_0 und B_1 der wahren Beziehung zwischen Y und X_1 sind unbekannt und sollen ermittelt werden. Diese Ermittlung erfolgt im Rahmen einer sogenannten **Schätzung dieses Modells**. Um eine Schätzung handelt es sich, da der wahre Einfluss der Kontakte auf die Verkaufsmengen aus den Daten nicht problemlos abzulesen ist. Dies erstens weil die vorhandenen Beobachtungen (die 15 Verkaufsregionen) nur eine **Stichprobe** (ein Sample) aus einer größeren **Grundgesamtheit** (Population) repräsentieren und zweitens die Absatzmengen noch von weiteren (zufälligen) Einflüssen abhängen, deren Gesamtwirkung sich im Fehlerterm E niederschlägt. Im PKW-Beispiel ist der wahre Zusammenhang der Gl. 2.2 ausnahmsweise bekannt, da es sich um einen konstruierten Datensatz handelt.[2]

Ausgangspunkt ist zur Veranschaulichung das Koordinatensystem in Abb. 2.2, das den grafischen Zusammenhang der Abb. 2.1 noch einmal wiederholt.

Eine Regressionsanalyse ermittelt, ob es einen **Zusammenhang zwischen Y und X_1** gibt, der sich in Form einer Gerade, d. h. als **lineare Funktion**, darstellen lässt. Es wird also eine zu den 15 Beobachtungen „passende" Gerade gesucht. Diese Gerade soll dem unbekannten wahren Zusammenhang entsprechen. Es stellt sich aber das Problem, anhand welches Kriteriums entschieden wird, ob die durchgehend **eingezeichnete Funktion „besser" oder „eher" passt** als eine andere lineare Funktion, bspw. die schwarz-gestrichelt eingezeichnete Funktion. Man könnte etwa folgendermaßen argumentieren: Die gestrichelte Gerade passt besser, weil sie genau durch vier Beobachtungspunkte läuft, die durchgehende Gerade im Gegensatz dazu durch keinen Beobachtungspunkt. Diese Überlegung ist aber wenig sinnvoll, denn es kommt ja nicht darauf an, einige wenige Beobachtungen sehr gut zu erklären, sondern es sollen möglichst alle Beobachtungen im Durchschnitt gut wiedergegeben werden. Auf diese Weise wird die unbekannte wahre Funktionsform der Gl. 2.2 am ehesten erfasst.

[2] Anhang 2.1 am Ende dieses Kapitels beschreibt die Konstruktion dieses Datensatzes und enthält damit auch die normalerweise unbekannten wahren Koeffizientenwerte B_0 und B_1.

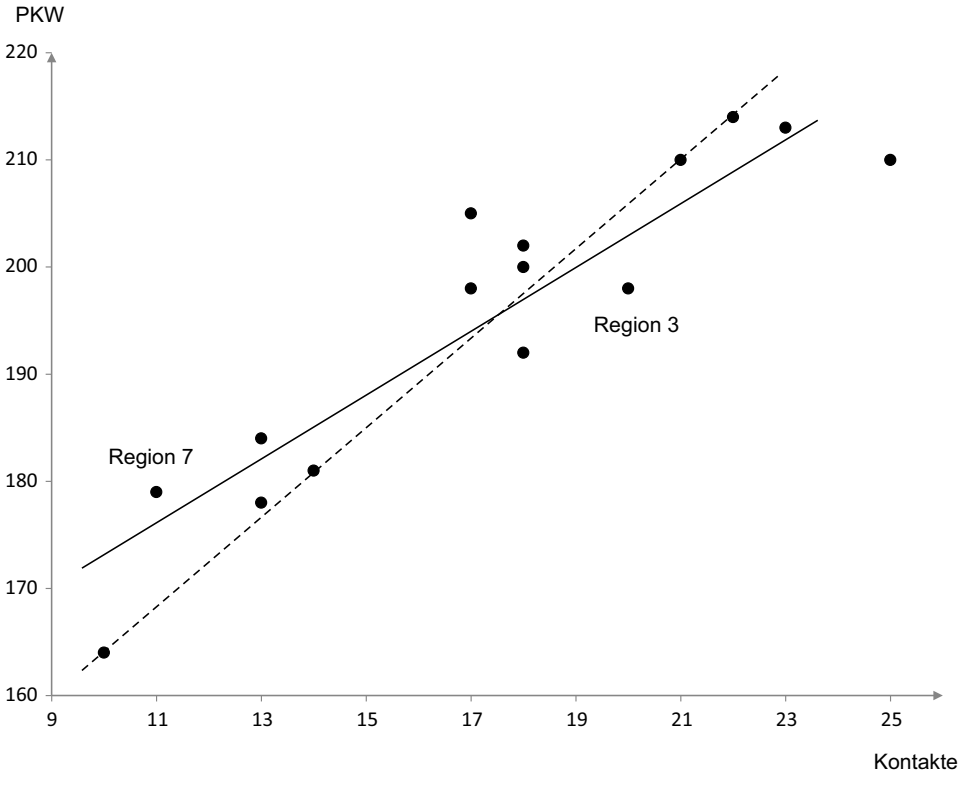

Abb. 2.2 Lineare Zusammenhänge I

Das Kriterium, das dazu benutzt wird, ist das Residuum (die Abweichung) e. Das Residuum e ist eine Schätzung des unbekannten wahren Fehlers E. Im Rahmen einer sogenannten **OLS-Schätzung** (Ordinary Least Squares = Methode der kleinsten Quadrate, abgekürzt **KQ-Methode**) wird die Funktion ermittelt, die dafür sorgt, dass die Summe der quadrierten Residuen minimiert wird. Das Residuum ist die Abweichung des Beobachtungswertes Y_i (siehe Abb. 2.3) vom entsprechenden Schätzwert $\hat{Y}_i$ (mit i = 1 bis 15).[3] Die geschätzten Werte werden mit einem Dach über der Variablen gekennzeichnet. Die beobachteten Werte haben dagegen kein Dach. Das Residuum ist bspw. für den Beobachtungspunkt der Region 7 mit 11 Kontakten und 179 verkauften PKW der vertikale Abstand

[3] Die Minimierung der Summe der Abweichungsquadrate ist ein sogenannter Schätzer (estimator), d. h. ein Verfahren, um eine passende Gerade zu ermitteln. Andere Schätzverfahren sind möglich (bspw. die Minimierung der Summe der absoluten Abweichungen). Die KQ-Methode hat im Vergleich erstens den Vorteil rechentechnisch einfach ermittelbar zu sein. Dies spielt heute aber dank der Leistungsfähigkeit der PCs keine Rolle mehr. Zweitens besitzt der KQ-Schätzer drei wünschenswerte Eigenschaften. Unter bestimmten Annahmen ist es ein unverzerrter, konsistenter und effizienter Schätzer (näheres dazu im Abschn. 5.1).

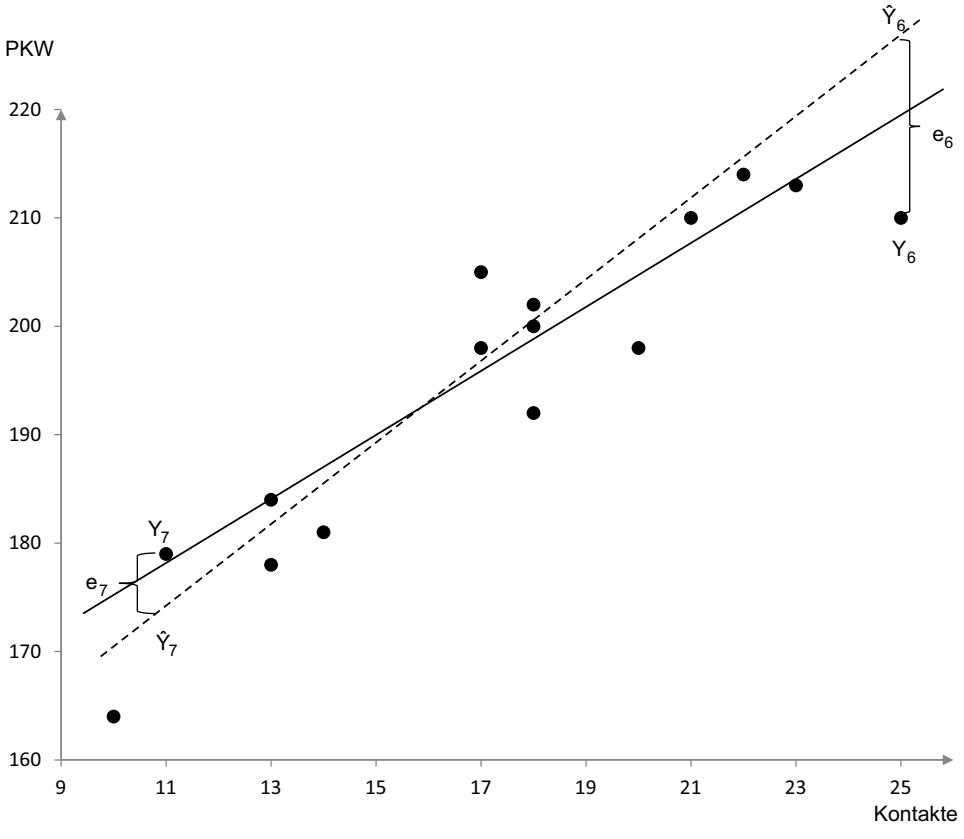

Abb. 2.3 Lineare Zusammenhänge II

zwischen dem Punkt 7 und der gestrichelt eingezeichneten Gerade:

$$e_7 = Y_7 - \hat{Y}_7 \quad \text{bzw.} \quad Y_7 = \hat{Y}_7 + e_7. \tag{2.3}$$

Entsprechend ist für die Region 6 das Residuum gleich e_6. Für die Region 7 handelt es sich um eine positive Abweichung und für die Region 6 um eine negative Abweichung. Die KQ-Methode quadriert alle Abweichungen und summiert diese. Sie ermittelt dann die Gerade, die diese Summe minimiert. Diese gestrichelte Gerade gibt die geschätzten Werte $\hat{Y}_i$ in Abhängigkeit von der Zahl der Kontakte wieder.

Aufgrund der Quadrierung behandelt die OLS-Methode positive und negative Abweichungen gleich und gewichtet „Ausreißer“ (also starke Abweichungen) besonders hoch. Dies führt im Beispiel dazu, dass die gestrichelt eingezeichnete Funktion der Abb. 2.3 bei einer OLS-Schätzung nicht als Ergebnis in Frage kommt, weil die Beobachtung der Region 6 mit 25 Kontakten und starker negativer Abweichung die Gerade „nach unten

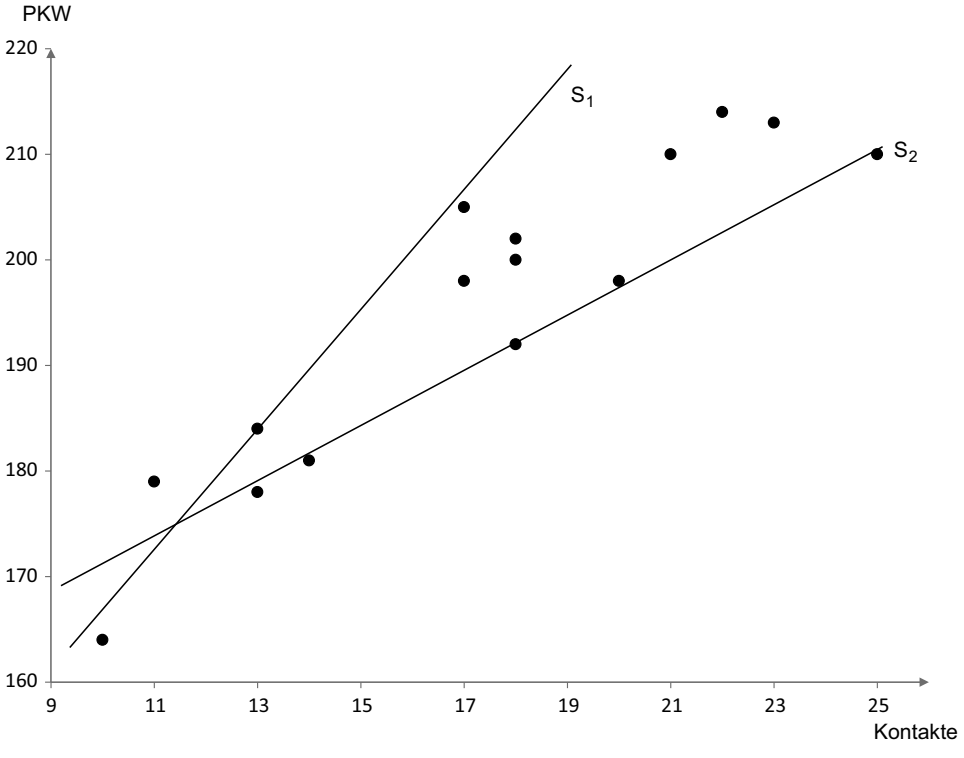

Abb. 2.4 Lineare Zusammenhänge III

zieht", d. h. flacher werden lässt. Durch die Quadrierung wird die Abweichung e_6 sehr groß, beeinflusst also die Steigung der Geraden sehr stark.

In Abb. 2.4 wird noch einmal deutlich, dass andere Schätzungen des Zusammenhangs von Y und X_1 denkbar sind, bspw. als Extreme die Schätzgeraden S_1 und S_2. Diese sind aber offensichtlich im Mittel aller Beobachtungen schlechtere Darstellungen des (vermutlichen) wahren Zusammenhangs.

Die Methode der kleinsten Quadrate berechnet also einen Wert für b_0 und einen Wert für b_1, die der durchgehend eingezeichneten Funktion in Abb. 2.2 entsprechen. Dies sind die sogenannten **Koeffizientenschätzungen** der unbekannten wahren Parameter B_0 und B_1. So ergeben sich für den Zusammenhang der Variablen *Menge* und *Kontakte* auf der Basis der Daten aus Tab. 1.2 die folgenden Werte (das Verfahren und dessen Rechenschritte werden im Abschn. 8.1 des Buchs erläutert):

$$\hat{Y} = 141{,}013 + 3{,}126X_1, \tag{2.4}$$

d. h. also $b_0 = 141{,}013$ und $b_1 = 3{,}126$.[4]

[4] Es werden hier und in den anschließenden Kapiteln die folgenden Abkürzungen verwendet: Die großen Buchstaben (B_0, B_1 usw) bezeichnen die uns unbekannten wahren Koeffizienten. Die kleinen

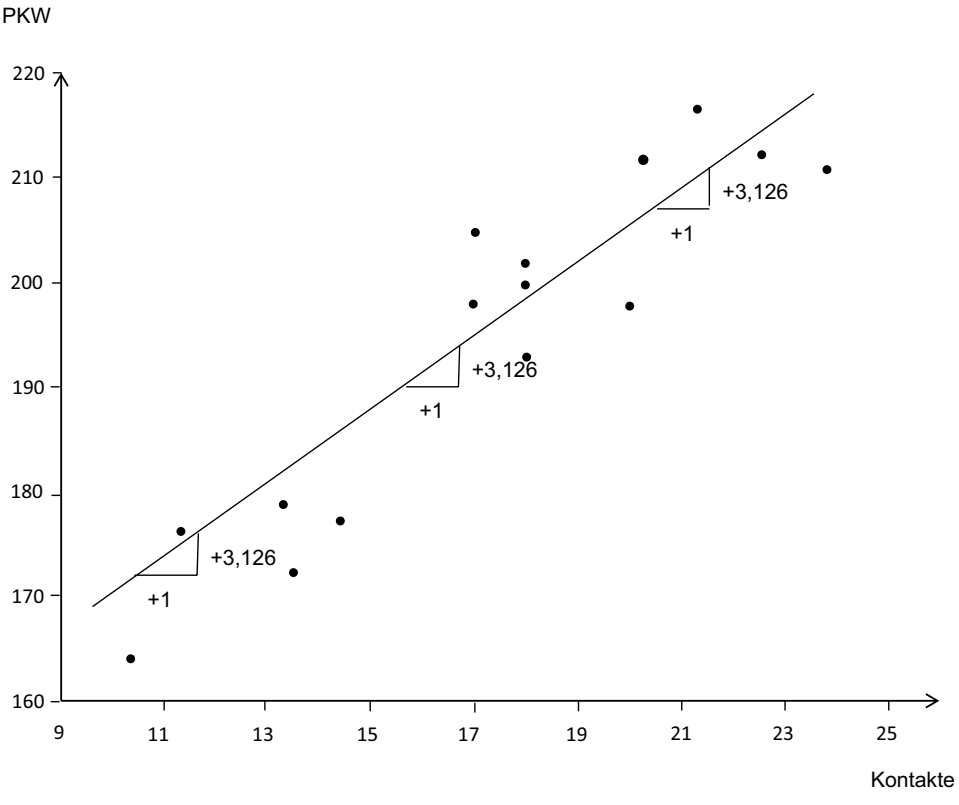

Abb. 2.5 Der marginale Effekt eines Kontaktes

Das Absolutglied (die **Konstante**)[5] b_0 ist der Schnittpunkt mit der Y-Achse und wie folgt zu interpretieren: Wenn die Zahl der Vertreterkontakte Null beträgt, liegt der Absatz im Durchschnitt aller Vertriebsregionen bei ca. 141 PKW pro Quartal. Bei b_1 handelt es sich um die **Steigung der Gerade**. Eine Erhöhung der Zahl der Kontakte um den Wert 1 lässt den PKW-Absatz im Durchschnitt um fast 3,13 Stück steigen. Diese Steigung ist mathematisch nichts anderes als die erste Ableitung der geschätzten Funktion nach X_1, also der **marginale Effekt** der Kontakte (Abb. 2.5).

Es wird in der Abbildung deutlich, dass der Effekt eines zusätzlichen Vertreterkontaktes wegen des linearen Zusammenhangs immer gleich groß ist – unabhängig davon, ob

Buchstaben (b_0, b_1 usw.) stehen für die geschätzten Koeffizienten. Sie werden in der Literatur häufig auch mit den griechischen Buchstaben (in der Regel β_0, β_1 usw.) abgekürzt. Zur Unterscheidung von wahren und geschätzten Parametern dienen hier und in der Literatur auch die „Dächer" über den Variablen bzw. Koeffizienten, so sind zum Beispiel $\hat{Y}$, $\hat{\beta}_1$ geschätzte Größen.

[5] Englisch: „Intercept" oder „Constant".

die Zahl von 11 auf 12 oder von 21 auf 22 Kontakte steigt, erhöht sich der PKW-Absatz um 3,126 Stück.

Durch Einsetzen konkreter Werte für die Zahl der Kontakte lässt sich berechnen, welche Absatzmengen zu erwarten sind. Bei 17 Vertreterkontakten ergibt sich eingesetzt in die geschätzte Gl. 2.4 folgendes:

$$\hat{Y} = 141{,}013 + 3{,}126 \times 17 = 194{,}16. \tag{2.5}$$

Die inhaltliche Interpretation ist, dass bei 17 Vertreterkontakten im Mittel 194,16 PKW verkauft werden. Der geschätzte Wert $\hat{Y}$ ist ein **bedingter Mittelwert** (**conditional mean**). Die Bedingung ist die Zahl der Vertreterkontakte, die stattfinden, im Beispiel also 17 Kontakte. Im Durchschnitt (im Mittel) aller Verkaufsregionen ist zu erwarten, dass bei 17 Kontakten 194,16 PKW verkauft werden.

Zur Wiederholung: Um in das Streudiagramm der Abb. 2.1 eine Gerade zu legen, die diese Beobachtungen möglichst „gut" wiedergibt, benötigen wir eine mathematische Regel, mittels derer sich b_0 und b_1 berechnen lassen. Das gängige Verfahren ist die KQ-Methode (OLS-Methode).[6] Diese Regeln zur Berechnung heißen Schätzer und das Ergebnis ist eine Schätzung der unbekannten Koeffizienten b_0 und b_1. Im Beispiel des OLS-Verfahrens betragen die Schätzungen für b_0 141,013 und b_1 3,126. Die Schätzung des Koeffizienten b_1 ist die erste Ableitung der Gl. 2.4, diese ist nichts anderes als die Steigung der Geraden und besagt inhaltlich, welche Auswirkungen ein zusätzlicher Vertreterkontakt auf die abhängige Variable, das heißt die Zahl der verkauften PKW ausübt.

Im Unterschied zu Korrelationskoeffizienten, die lediglich einen dimensionslosen Zusammenhang beschreiben, gehen die Regressionskoeffizienten von einer eindeutigen Wirkungsrichtung (Kausalrichtung) aus: Die Zahl der Kontakte determiniert die Absatzmenge.[7]

Bei der Interpretation wird unterstellt, dass die **Residuen e** lediglich **Zufallseinflüsse** wiedergeben (bspw. Messfehler bei der Datenerhebung u. ä.) und der „wahre" Zusammenhang durch die geschätzte Funktion erfasst wird. Die Wahrheit ist hier als der im Mittel zu erwartende Wert der abhängigen Variablen definiert. Allerdings sind noch einige Aspekte zu prüfen, bevor an diese „Wahrheit" geglaubt wird (siehe Kap. 5 bis 7).

Im nächsten Schritt ergibt sich für den Vertriebsverantwortlichen die Frage, welche Relevanz und welchen Einfluss die Preispolitik auf den Absatz hat, denn im geplanten Meeting mit dem Vorstandsverantwortlichen für das Marketing in Europa soll die Möglichkeit der aufgrund von Kostensteigerungen notwendigen Preiserhöhungen auch auf dem deutschen Markt besprochen werden.

---

[6] Es existieren auch andere Regeln zur Berechnung, die im Anhang 5.1 des Kap. 5 kurz beschrieben werden. Die OLS-Methode ist aber das Referenzverfahren und grafisch besonders eingängig vermittelbar.

[7] Zum Unterschied von Regressionskoeffizient und Pearson-Korrelationskoeffizient siehe auch Abschn. 8.1 und 8.2 am Ende des Buchs.

Analog zum Vorgehen bei der einfachen Analyse der Vertreterkontakte enthält Abb. 2.6 eine grafische Widergabe der Verkaufszahlen und Preise in den 15 Vertriebsregionen. Das Ergebnis der dazugehörenden einfachen Regressionsanalyse lautet:

$$\hat{Y} = 149{,}098 + 3{,}196X_2. \tag{2.6}$$

Abb. 2.6 und die dazu mittels OLS ermittelten Koeffizientenschätzungen offenbaren aber im Vergleich zum Einfluss der Vertreterkontakte eine Reihe von Schwierigkeiten. Die Variable *Preis* hat einen positiven Einfluss, da der Koeffizientenwert +3,196 beträgt. Inhaltlich führt damit eine Preiserhöhung um 1000 € zu einer Steigerung des Absatzes um im Durchschnitt ca. 3,2 PKW. Der Zusammenhang ist nach Abb. 2.6 allerdings weniger eindeutig und unter Umständen für höhere Preise (über 15.500 €) sogar falsch. Dass Preiserhöhungen auf dem hart umkämpften Markt für PKW der untersten Mittelklasse eine verkaufsförderliche Maßnahme darstellen, ist ökonomisch kaum sinnvoll und ein Indiz dafür, dass bei der Analyse etwas Grundlegendes übersehen worden ist.

Tatsächlich wird bei dieser isolierten Betrachtung des Einflusses der Preise vernachlässigt, dass es ja noch weitere wichtige Einflussfaktoren gibt. Im PKW-Beispiel sind

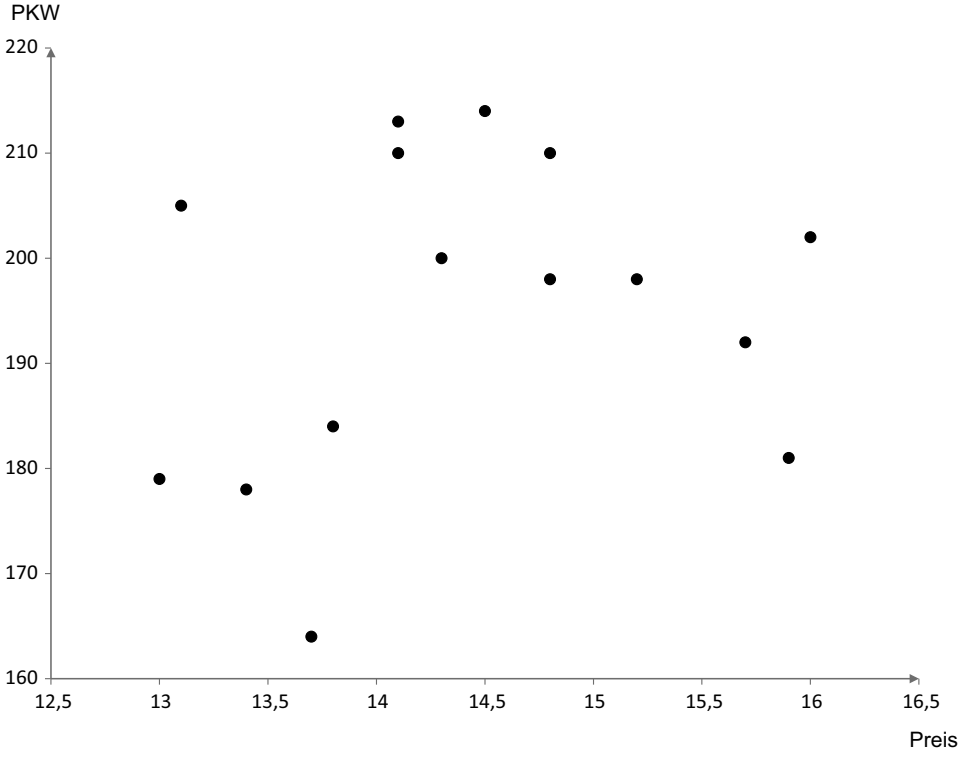

Abb. 2.6 Einfluss des Preises

http://www.springer.com/978-3-662-53823-4

Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 1
Eine nichtmathematische Einführung mit SPSS und Stata
Stoetzer, M.-W.
2017, XII, 326 S. 133 Abb., Softcover
ISBN: 978-3-662-53823-4