Einfache lineare Regression

2

Wie bereits erwähnt, besteht das Hauptziel dieses *essentials* darin, die Basis der Regressionsmodelle zu vermitteln. Deshalb berücksichtigen wir in der Modellbildung lediglich eine einzige Einflussgröße. Zudem setzen wir voraus, dass sowohl die Einfluss- als auch die Zielvariable metrisch sind. Eine Modellierung von Fällen, in denen mehrere Einflussgrößen vorkommen, die Daten binär oder kategorial sind, baut in der Regel auf diesem Grundmodell auf. Das Buch *Regression* der Autoren Fahrmeir et al. (2009) gibt einen Überblick darüber.

Das Streudiagramm

Bevor ein Modell ausgewählt wird, sollte man sich ein grobes Bild von den Daten machen. Dadurch werden die ersten Gründe dafür gelegt, warum – in unserem Fall – ein linearer Ansatz infrage kommt. Einen ersten Eindruck von den Daten gewinnen wir, indem wir die erhobenen Datenpaare $(x_i, y_i),\ i = 1, 2, \ldots, n$ als Punkte auf einer x-y-Ebene darstellen. Wir erhalten ein *Streudiagramm* oder eine *Punktewolke*. Wenn die „Wolke" uns an eine schräg liegende Ellipse erinnert, können wir versuchen, die Daten durch eine lineare Funktion zu modellieren. Das bedeutet: Wir gehen von der Vorstellung aus, dass es zwischen den Variablen x und y in Wahrheit eine Abhängigkeit der Art $f(x) = a + bx$ mit $a, b \in \mathbb{R}$ gibt, welche aber durch nicht systematische oder Zufallsfehler überlagert wird, sodass nicht $(x_i, f(x_i))$ erscheinen, sondern (x_i, y_i), wobei $y_i = f(x_i) + Fehler$. Wir nehmen also an, dass Zufallsfehler dafür verantwortlich sind, dass die Werte nicht auf einer Geraden liegen, sondern um sie streuen.

Als Beispiel betrachten wir das Streudiagramm der bereits erwähnten Körpergrößen-Daten von $n = 400$ Vater-Sohn-Paaren (Keller 2009, Exercise 17.2,

I. Frost, *Einfache lineare Regression*, essentials,
https://doi.org/10.1007/978-3-658-19732-2_2

S. 550) in Abb. 2.1. Auf der x-Achse werden die Körpergrößen der Väter eingetragen, auf der y-Achse die der Söhne. Jeder einzelne Punkt (x_i, y_i) auf dem Streudiagramm vertritt die Körpergröße x_i des Vaters und die seines Sohnes y_i (in Inches). Die Punktewolke hat nahezu Ellipsenform mit einer positiven Neigung.

Deutlich zu erkennen ist der isolierte Punkt am unteren linken Rand der Ellipse. Die Ursache für einen solchen Ausreißer kann ein Mess- bzw. Übertragungsfehler sein; er kann aber auch einen tatsächlich vorgekommenden Wert darstellen. Um das festzustellen ist es üblich und angebracht, diesen Wert genauer unter die Lupe zu nehmen. Aus Platzgründen verfolgen wir dieses Thema hier nicht weiter und schließen diesen Wert einfach in die Berechnung mit ein.

Mit der Regressionsanalyse können wir die grundlegende wahre Funktion sicherlich nicht finden. Wir können jedoch mittels der von C.F. Gauß (1777–1855) entwickelten Methode der kleinsten Quadrate eine lineare Funktion finden, die die Beziehung zwischen diesen Variablen *am besten* beschreibt (siehe Kap. 3). Besitzen die Zufallsfehler gewisse stochastische Eigenschaften, sind Gütekriterien für das Modell möglich. Welche Eigenschaften die Zufallsfehler haben sollen, zeigt nachfolgender Abschnitt.

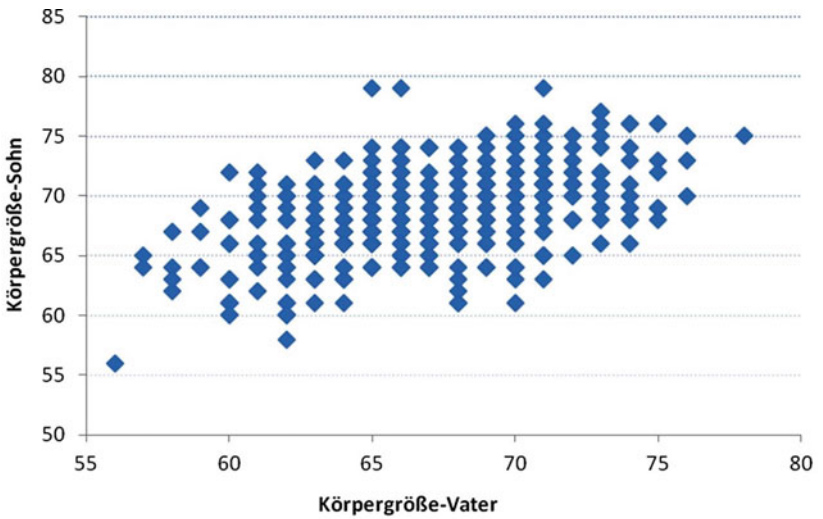

Abb. 2.1 Das Streudiagramm der Körpergrößen von Vater-Sohn-Paaren (in Inches)

Das lineare Regressionsmodell

Mit einem Modell wollen wir herausfinden, wie eine *Einflussvariable* x das Eintreten einer *Zielvariablen*[1] y erklären kann. Unter der Annahme, dass y von x linear abhängig ist, wird der Zusammenhang zwischen den Variablen durch eine lineare Funktion modelliert. Der Zusammenhang ist nicht deterministisch, sondern durch zufällige Fehler additiv überlagert; wir schreiben $y = a + bx + u,\ a, b \in \mathbb{R}$. Anders ausgedrückt, gilt für jede Beobachtung i die Modellgleichung

$$y_i = a + bx_i + u_i,\ a, b \in \mathbb{R},\ i = 1, 2, \ldots, n.$$

Die Fehlervariablen u_i umfassen alle unsystematischen, zufälligen Fehler. Für $u_i,\ i = 1, 2, \ldots, n$, setzen wir voraus:

1. $E(u_i) = 0$
2. $Cov(u_i, u_j) = 0$ für $i \neq j$
3. $Var(u_i) = \sigma^2 < \infty$
4. Für die statistische Inferenz verlangt man zusätzlich, dass u_i normalverteilt sind mit $E(u_i) = 0$ und $Var(u_i) = \sigma^2$.

Sehen wir uns die einzelnen Voraussetzungen genauer an.

Der Erwartungswert[2] *von u_i ist gleich Null* in der Annahme 1 bedeutet, dass die Wirkung der Fehlervariablen sich im Mittel aufheben. Dies ist eine plausible

[1] Andere Bezeichnungen für Einflussvariablen sind Regressoren, unabhängige, erklärende oder exogene Variablen. Für die Zielvariable findet man alternative Bezeichnungen wie Regressand, abhängige, erklärte oder endogene Variable.

[2] Der Begriff *Erwartungswert* wurde von dem holländischen Physiker Christian Huygens (1629–1695) im Zusammenhang mit Glücksspielen eingeführt. Den erwarteten Gewinn (oder Verlust) eines Glücksspieles kann ein Spieler berechnen, indem er die Wahrscheinlichkeit eines jeden Spielausgangs mit dem Geldbetrag, den er gewinnen (oder verlieren) kann, multipliziert und anschließend alle Ergebnisse aufsummiert. Später wird der Begriff allgemeiner für Zufallsvariablen definiert. Sehr vereinfacht kann man sich den Erwartungswert einer Zufallsvariablen als den Mittelwert vorstellen, den die Zufallsvariable auf lange Sicht annehmen kann. Anzumerken ist jedoch, dass es Zufallsvariablen gibt, deren Erwartungswert nicht existiert. Das berühmteste Beispiel hierfür ist das von Nikolaus Bernoulli (1687–1759) vorgestellte Sankt Petersburger Spiel, bei dem ein Spieler eine unverfälschte Münze solange wirft, bis *Kopf* zum ersten Mal erscheint. Das Erscheinen von *Kopf* zum ersten Mal beendet auch das Spiel. Erscheint beim ersten Wurf *Kopf*, erhält der Spieler 2 Rubel (und das Spiel ist vorbei). Wenn das Ergebnis des ersten Wurfes *Zahl* und das des zweiten *Kopf* ist, erhält der Spieler 4 Rubel. Erscheint *Kopf* zum ersten Mal beim dritten Wurf, kann der Spieler mit einem Gewinn von 8 Rubel nach Hause gehen. Bei jedem weiteren Wurf verdoppelt sich also der Gewinn. Diese

Annahme für zufällige Fehler (zufällige Fehler sollen ja keinen systematischen Effekt mehr enthalten).

Die Kovarianz von zwei unterschiedlichen Störgrößen u_i und u_j ist gleich Null ist der Inhalt der zweiten Voraussetzung. Wegen $E(u_i) = 0$ für alle $i = 1, 2, \ldots, n$ ist eine Kovarianz gleicht Null gleichbedeutend mit der Unkorreliertheit der Variablen. Liegt eine Normalverteilung vor, bedeutet dies eine Unabhängigkeit der Störvariablen (zu *Korrelation* und *Unabhängigkeit* siehe Kap. 6). Korrelierte Störgrößen bedeuten, dass Abweichungen von der linearen Funktion nicht mehr zufällig sind. Eine Beobachtung wäre dann beispielsweise von der vorangegangenen abhängig.

Die dritte Voraussetzung garantiert die Existenz der Varianzen sowie ihre Eigenschaft, für alle $i = 1, 2, \ldots, n$ konstant gleich σ^2 zu sein. Diese Eigenschaft nennt man *Homoskedastizität*. Sind die Varianzen nicht konstant, heißen sie *heteroskedastisch* (Heteroskedastzität kommt insbesondere in Zeitreihendaten vor). Unter diesen Annahmen gilt für ein festes x:

$$E(y) = a + bx \quad \text{und} \quad Var(y) = \sigma^2$$

Die Gerade $a + bx$ modelliert somit den Erwartungswert der Zielvariablen y, also stellt sie eine Mittelgerade für y dar. Insbesondere ist y unter der Normalverteilungsannahme auch normalverteilt. Damit können wir Konfidenzintervalle bilden und statistische Tests durchführen.

In der oben angegebenen Modellgleichung sind die Koeffizienten a, b, die Fehlervariable u_i sowie die Varianz σ^2 unbekannt. Wir werden sie aus den Daten (x_i, y_i) schätzen. Davor geben wir die wichtigsten Datenkennzahlen an.

Kennzahlen zur Beschreibung von Daten

Die wichtigsten Kennzahlen sind das arithmetische Mittel und die Standardabweichung. Das arithmetische Mittel ist die Datensumme geteilt durch deren Anzahl. Wenn wir im Alltag von einem Durchschnitt sprechen, meinen wir eben das arithmetische Mittel[3]. Die Standardabweichung ist definiert als die Wurzel der mittleren quadratischen Abweichung der Daten von ihrem arithmetischen Mittel und kann als

(Fortsetzung 2 continued)
Spielregel führt zu einem unendlich hohen Gewinn. Denn der Erwartungswert des Gewinns ergibt sich gemäß: $E(G) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \ldots = 1 + 1 + 1 + \ldots = \infty$.

[3] In der Datenanalyse existieren weitere Kennzahlen, die Durchschnittswerte darstellen, etwa der Modalwert (die häufigste Beobachtung) oder der Median (der die Daten in zwei Hälften teilt).

eine Maßzahl für die durchschnittliche Streuung der Daten um ihren arithmetischen Mittelwert angesehen werden.

Wenn wir die Daten mit $x_1, x_2, \ldots, x_n$ bezeichnen, symbolisiert traditionsgemäß $\bar{x}$ das arithmetische Mittel, s_x^2 die mittlere quadratische Abweichung und $s_x = \sqrt{s_x^2}$ die Standardabweichung der Variablen x (entsprechend für y). Formal berechnen wir $\bar{x}$ sowie s_x^2 gemäß:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Zudem geben der größte und der kleinste Wert die Spannweite der Beobachtungen an (auf mögliche Ausreißer achten!). Nützlich sind auch der Median, der eine aufsteigend geordnete Datenreihe in zwei Hälften aufteilt sowie das untere und das obere Quartil, Q_1 und Q_3. Die Quartile werden ebenso für eine aufsteigend geordnete Datenreihe bestimmt. Ein Viertel der Daten liegen unterhalb und drei Viertel oberhalb von Q_1. Das obere Quartil teilt umgekehrt die Daten in drei Viertel unterhalb und in ein Viertel oberhalb von Q_3 auf.

Im Streudiagramm markiert der Punkt $(\bar{x}, \bar{y})$ das Zentrum der Daten (x_i, y_i), $i = 1, 2, \ldots, n$. Die Streuung um dieses Zentrum gibt die Kovarianz

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

an. Je heterogener die Werte sind, desto größer wird betragsmäßig die Kovarianz (im Gegensatz zu s^2, das niemals negativ wird, kann s_{xy} positiv oder negativ sein).

Für unsere Beispieldaten zeigt Tab. 2.1 die zugehörigen Kennzahlen (gerundet).

Tab. 2.1 Kennzahlen der Körpergrößen (Inches) der Väter bzw. der Söhne

Körpergröße	Min	Q_1	Median	Q_3	Max	Arithm. Mittel	Standardabweichung
Vater	56	64	67	70	78	67,14	4,05
Söhne	56	66	69	71	79	68,70	3,76

All diese Werte zeigen, dass beide Gruppen sehr ähnlich strukturiert sind; insbesondere erkennen wir, dass die Väter und die Söhne im Mittel nahezu gleich groß mit etwa gleicher Standardabweichung sind (im Mittel sind die Väter 67,14 und die Söhne 68,70 Inches groß; die Standardabweichungen betragen 4,05 bzw. 3,76 Inches; auch sind die anderen Kennzahlen ähnlich groß). Die Kovarianz ist $s_{xy} = 7{,}87$.

http://www.springer.com/978-3-658-19731-5

Einfache lineare Regression
Die Grundlage für komplexe Regressionsmodelle verstehen
Frost, I.
2018, VIII, 37 S. 8 Abb., Softcover
ISBN: 978-3-658-19731-5