8: Lineare Regression

Statistische Verfahren in der Geographie

Till Straube <straube@geo.uni-frankfurt.de>

Institut für Humangeographie
Goethe-Universität Frankfurt

1 Lernziele dieser Sitzung

Sie können…

- eine Regressionsgerade berechnen.
- Werte aus der Regressionsgerade ableiten.
- Residuen errechnen.
- den Determinationskoeffizienten R^2 berechnen und interpretieren.

2 Regresssionsanalyse

> Sind zwei stochastisch abhängige Variablen x und y durch eine Regressionsgleichung miteinander verknüpft, kann die eine Variable zur Vorhersage der anderen eingesetzt werden. (Bortz und Schuster 2010: 183)

Es gibt viele Möglichkeiten, Regressionen zu modellieren. Im Rahmen dieser Veranstaltung wird nur die lineare Regression (engl. *linear regression*) behandelt. Lineare Regressionsmodelle werden immer durch eine lineare Gleichung des Formats

$$y = a + b \cdot x \tag{1}$$

ausgedrückt, wobei a der Achsenabschnitt ist und b die Steigung. Ist die Gleichung bekannt, so können wir für jeden Wert x einen entsprechenden Wert y „vorhersagen".

Abbildung 1 zeigt ein solches lineares Regressionsmodell als Gerade durch ein Streudiagramm.

Der Achsenabschnitt $a \approx 2{,}2$ bedeutet, dass die Regressionsgerade die y-Achse etwa auf der Höhe 2,2 schneidet (bei $x = 0$). Die Steigung $b \approx 1{,}7$ heißt, dass für jede zusätzliche Einheit der Variable x ca. 1,7 zusätzliche Einheiten der Variable y erwartet werden können.

Wenn die Regressionsgleichung bekannt ist, kann für jedes gültige (grundsätzlich: jedes beliebige) x ein erwarteter Wert $\hat{y}$ berechnet werden. So könnte uns bei der Beispielregression interessieren, welchen Wert $\hat{y}_i$ im Modell annimmt, wenn $x_i = 20$ beträgt:

$$\begin{aligned} \hat{y}_i &= a + b \cdot x_i \\ &\approx 2{,}2 + 1{,}7 \cdot 20 \\ &= 36{,}2 \end{aligned}$$

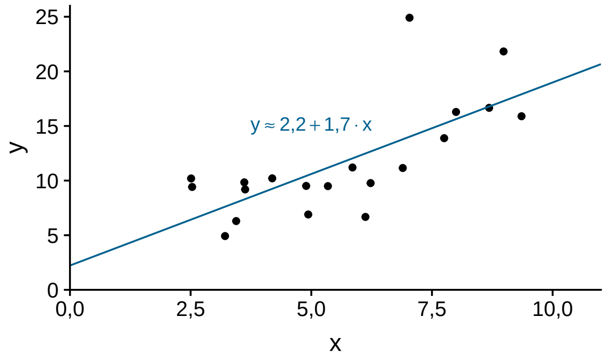

Abbildung 1: Regressionslinie durch ein Streudiagramm

Bei solchen Schätzungen *außerhalb* des bekannten Wertebereichs spricht man auch vom „Extrapolieren“, sonst – für fehlende Werte innerhalb des bekannten Wertebereich – vom „Interpolieren“.

Umgekehrt könnte die Frage lauten: Wie groß muss ein x_i sein, damit (im Modell) $\hat{y}_i = 12$ beträgt? Dies lässt sich durch eine einfache Umformung der Gleichung 1 berechnen:

$$\begin{aligned}
\hat{y}_i &= a + b \cdot x_i \\
x_i &= \frac{\hat{y}_i - a}{b} \\
&= \frac{12 - 2{,}2}{1{,}7} \\
&\approx 5{,}8
\end{aligned}$$

Bei der Regressionsanalyse wird ein gerichtetes Abhängigkeitsverhältnis der Variablen impliziert: y hängt hier von x ab. Daher wird x auch die „Prädiktorvariable“ und y die „Kriteriumsvariable“ genannt.

Softwarehinweis

Wenn in R ein lineares Modell (eine Regressionsgerade) vorliegt, können Werte mit `predict()` geschätzt werden.

Es ist also für derartige Fragestellungen nötig, die Gleichung der Regressionsgeraden zu kennen. Im Folgenden wird gezeigt, wie diese anhand einer bivariaten Verteilung bestimmt werden kann.

3 Bestimmung der Regressionsgeraden

Der Koeffizient b (also die Steigung der Regressionsgeraden) lässt sich berechnen, indem man die Kovarianz s_{xy} durch die Varianz von x dividiert:

$$b = \frac{s_{xy}}{s_x^2} \tag{2}$$

Tabelle 1: Messwerte am Frankfurter Flughafen

Aufenthaltszeit (min)	Ausgaben (€)
x_i	y_i
121	17,94
125	23,15
293	44,31
370	42,46
246	35,51
281	28,46
169	18,47
328	56,77
388	40,11
131	12,64
299	24,54
324	46,37

Der Koeffizient a (also der Achsenabschnitt) ergibt sich wiederum aus b und den Mittelwerten $\bar{x}$ und $\bar{y}$:

$$a = \bar{y} - b \cdot \bar{x} \tag{3}$$

Softwarehinweis

In R lässt sich ein lineares Regressionsmodell mit dem Befehl `lm()` erstellen.

Die Bestimmung der Regressionsgeraden soll nun mit einem Beispiel illustriert werden.

3.1 Beispiel

Wir fragen uns, wie die Aufenthaltszeit von Passagieren am Frankfurter Flughafen mit dem Betrag zusammenhängt, den sie in den dortigen Geschäften ausgeben. Eine Zufallserhebung habe die Werte in Tabelle 1 ergeben.

Mit den Methoden aus Sitzung 2 und 7 können wir folgende Werte für die Mittelwerte $\bar{x}$ und $\bar{y}$, die Varianz s_x^2 sowie die Kovarianz s_{xy} berechnen:

$$\begin{aligned} \bar{x} &= 256{,}25 \\ \bar{y} &\approx 32{,}56 \\ s_x^2 &\approx 9340{,}93 \\ s_{xy} &\approx 1062{,}50 \end{aligned}$$

Für die Steigung der Regressionsgeraden b setzen wir die entsprechenden Werte in Gleichung 2 ein:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ &\approx \frac{1062{,}50}{9340{,}93} \\ &\approx 0{,}114 \end{aligned}$$

Die Steigung von 0,114 bedeutet, dass – im linearen Regressionsmodell – Passagiere in jeder zusätzlichen Minute, die sie am Flughafen verbringen, in etwa 11,4 zusätzliche Cent ausgeben.

Der Achsenabschnitt a berechnet sich dann gemäß Gleichung 3:

$$\begin{aligned} a &= \bar{y} - b \cdot \bar{x} \\ &\approx 32{,}56 - 0{,}114 \cdot 256{,}25 \\ &\approx 3{,}35 \end{aligned}$$

Dieser Wert ergibt nur einen abstrakt-mathematischen Sinn – es dürfte in der Praxis wohl kaum Passagiere geben, die 0 Minuten am Flughafen verbringen und € 3,35 ausgeben.

Mit dem Achsenabschnitt a und der Steigung b lässt sich folgende Gleichung für die Regressionsgerade aufstellen (s. Gleichung 1):

$$\begin{aligned} y &= a + b \cdot x \\ y &\approx 3{,}35 + 0{,}114 \cdot x \end{aligned}$$

Graphisch ist diese lineare Regression in Abbildung 2 dargestellt.

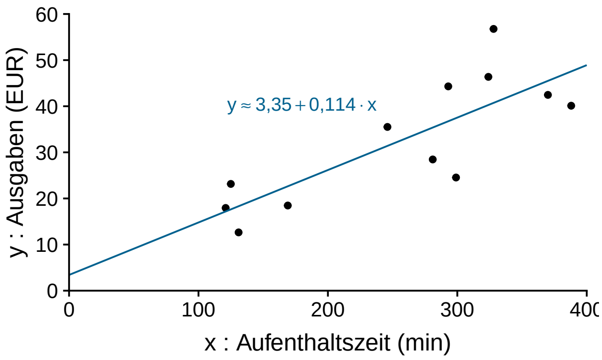

Abbildung 2: Regressionslinie durch ein Streudiagramm

4 Residuen

Residuen (engl. *residuals*) werden mit e bezeichnet und sind die Differenzen zwischen den tatsächlichen y-Werten und den im Modell erwarteten $\hat{y}$-Werten für die jeweiligen x-Werte:

$$e_i = y_i - \hat{y}_i \tag{4}$$

Tabelle 2: Residuen der Beispielwerte

Aufenthaltszeit (min)	Ausgaben (€)	Erwartete Ausgaben (€)	Residuen (€)
x_i	y_i	$\hat{y}_i \approx 3{,}35 + 0{,}114 \cdot x_i$	$e_i = y_i - \hat{y}_i$
121	17,94	17,144	0,796
125	23,15	17,600	5,550
293	44,31	36,752	7,558
370	42,46	45,530	-3,070
246	35,51	31,394	4,116
281	28,46	35,384	-6,924
169	18,47	22,616	-4,146
328	56,77	40,742	16,028
388	40,11	47,582	-7,472
131	12,64	18,284	-5,644
299	24,54	37,436	-12,896
324	46,37	40,286	6,084

Residuen sind also – auch dem Wortstamm nach – das, was nach der Vorhersage durch das Modell „übrig bleibt" von den tatsächlich beobachteten Werten (also der Teil des Werts, der *nicht* durch das Regressionsmodell erklärt wird).

Softwarehinweis

Residuen lassen sich in R durch den Befehl `resid()` errechnen.

4.1 Beispiel

Graphisch sind die Residuen für unser Beispiel in Abbildung 3 dargestellt (positive Werte in grün, negative Werte in rot), tabellarisch in Tabelle 2.

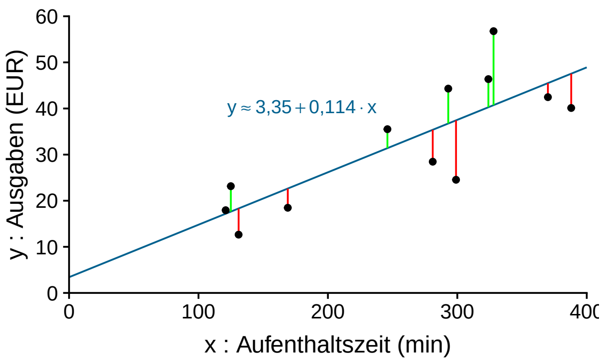

Abbildung 3: Graphische Darstellung der Residuen

Residuen spielen in vielen statistischen Verfahren eine Rolle, z.B. in der Residuenanalyse. Diese Verfahren werden im Rahmen dieser Veranstaltung jedoch nicht behandelt.

5 Determinationskoeffizient

Der Determinationskoeffizient R^2 (engl. *coefficient of determination*) ist formal definiert als das Verhältnis der Varianz der vorhergesagten $\hat{y}$-Werte zur Varianz der tatsächlich beobachteten y-Werte (wobei sich der Term $[n-1]$ auskürzt):

$$R^2 = \frac{\sum\limits_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

Da Zähler und Nenner als Quadratsummen stets positiv sind und die Varianz der $\hat{y}$-Werte immer *kleiner oder gleich* der Varianz der y-Werte ist, nimmt der Determinationskoeffizient immer einen Wert zwischen 0 und 1 an.

Je größer R^2, desto besser erklärt das lineare Regressionsmodell die tatsächlich beobachteten Werte. $R^2 = 1$ bedeutet, dass das Modell die Werte perfekt erklärt.

Für lineare Regressionsmodelle (also für die einzige Regression, die im Rahmen dieser Veranstaltung behandelt wird) lässt sich R^2 auch berechnen, indem wir den Korrelationskoeffizienten r quadrieren:

$$R^2 = r^2 \tag{6}$$

Softwarehinweis

In R wird mit dem Befehl `summary()` unter anderem der Determinationskoeffizient eines linearen Regressionsmodells ausgegeben.

5.1 Beispiel

Mit den Methoden aus Sitzung 7 können wir den Korrelationskoeffizienten für unser Beispiel errechnen:

$$\begin{aligned} r &= \frac{s_{xy}}{s_x \cdot s_y} \\ &\approx \frac{1062{,}50}{96{,}65 \cdot 13{,}68} \\ &\approx 0{,}804 \end{aligned}$$

Der Determinationskoeffizient ergibt sich dann mit Gleichung 6:

$$\begin{aligned} R^2 &= r^2 \\ &\approx 0{,}804^2 \\ &\approx 0{,}646 \end{aligned}$$

6 Aufgaben

6.1 Aufgabe 1

Sie haben für eine bivariate Verteilung die folgende Regressionsgleichung bestimmt:

$$y = -1{,}48 - 0{,}975 \cdot x$$

a) Bestimmen Sie die erwarteten $\hat{y}_i$-Werte für diese x_i-Werte:

$$0{,}3 \quad -18{,}5 \quad -13{,}5 \quad -17{,}2 \quad 29{,}8 \quad 25{,}6 \quad -36{,}4 \quad -26{,}2$$

b) Für welche Werte x_i sagt das Regressionsmodell diese Werte $\hat{y}_i$ voraus?

$$-10 \quad 15 \quad -50 \quad -10 \quad -60 \quad -55 \quad -20 \quad 0$$

c) Bestimmen Sie die Residuen für die tatsächlich beobachtete Messreihe:

x_i	y_i
-11,49	6,82
8,22	-8,59
-25,66	25,92
23,81	-26,91
-3,14	4,41
-1,52	-3,39
20,15	-19,89
-10,22	9,30

6.2 Aufgabe 2

Eine bivariate Verteilung sei gekennzeichnet durch die folgenden Parameter:

$$\bar{x} = 157{,}5$$
$$\bar{y} = 156{,}7$$
$$s_x^2 = 1080{,}94$$
$$s_y^2 = 884{,}46$$
$$s_{xy} = 869{,}83$$

a) Bestimmen Sie die Regressionsgleichung im linearen Modell.

b) Bestimmen Sie den Determinationskoeffizienten R^2.

6.3 Aufgabe 3

Diese Aufgabe erfordert auch Verfahren aus Sitzung 6.

Sie fragen sich, wie die erreichte Punktzahl in einer Klausur mit der Vorbereitungszeit der geprüften Studierenden zusammenhängt. Sie erheben die folgende Messreihe:

Vorbereitungszeit (min)	Erreichte Punktzahl
834	88
17	41
519	75
253	39
739	77
844	100

a) Welche Punktzahl ist mit einer Vorbereitunszeit von sechs Stunden zu erwarten?

b) Ab welcher Vorbereitungszeit ist im Modell zu erwarten, dass ein Studierende*r die Klausur besteht ($\geq$ 50 Punkte)?

c) Ab welcher Vorbereitungszeit kann laut Modell mit der vollen Punktzahl (100 Punkte) gerechnet werden?

d) Wie gut erklärt ein lineares Modell die Prüfungsleistungen anhand der Vorbereitungszeit?

e) Welche Limitationen hat das Modell? Denken Sie an extreme Werte.

7 Tipps zur Vertiefung

- Kapitel 11 in Bortz und Schuster (2010)
- Kapitel 6.2 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 17 in Klemm (2002)

Quellen

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. 5. Aufl. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Klemm, Elmar. 2002. *Einführung in die Statistik. Für die Sozialwissenschaften*. Wiesbaden: Westdeutscher Verlag.