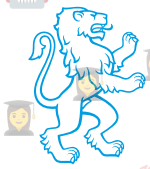


Kantonsschule Im Lee

Informatik

Aus Daten Lernen

Skript

Cyril Wendl

Inhaltsverzeichnis

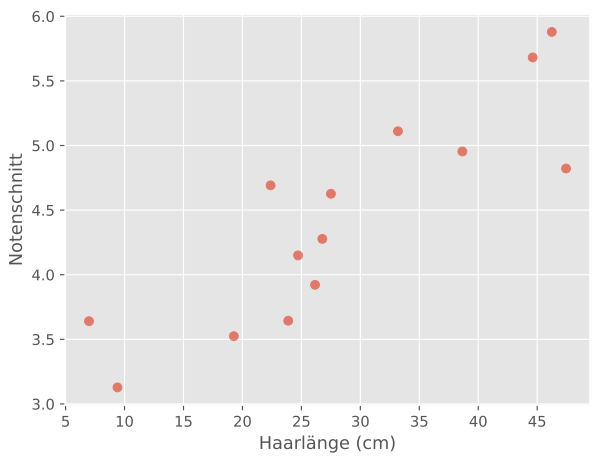
Kapitel 1

Daten vorhersagen

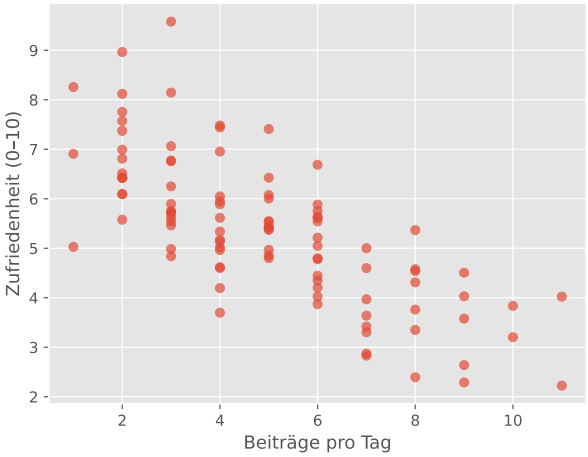
Eine der häufigsten Anwendungen in der künstlichen Intelligenz ist die Vorhersage von Daten. Dabei wird ein Modell erstellt, das auf Basis von vorhandenen Daten Vorhersagen für zukünftige Daten trifft.

Bevor wir uns mit der Vorhersage von Daten befassen, sollten wir zuerst einige grundlegende Begriffe klären.

?? zeigt zwei Beispiele für Datenreihen (Variablen). Die linke Grafik zeigt eine Datenreihe der Verteilung der Noten sowie der Haarlänge in einer weiblichen Schulklasse. Die rechte Grafik zeigt eine Datenreihe der Nutzung von Social Media und der allgemeinen Lebenszufriedenheit der BenutzerInnen.



(a) Haarlänge und Notenschnitt in einer Schulklasse



(b) Social-Media-Nutzung und Zufriedenheit
ig:socialmedia$_v s_z ufriedenheitub@fig$:
$socialmedia_v s_z ufriedenheit$

Abbildung 1.1: Beispiele für Korrelationen
ig:Korrelationen

Die linke Grafik zeigt einen positiven Zusammenhang zwischen der Haarlänge und der Schulnote. Bei der Interpretation solcher Daten ist Vorsicht geboten: Es könnte sein, dass die Daten rein zufällig so aussehen. Wenn man mehrere Klassen anschaut, sähe man eventuell keinen Trend. Eine allfällige These, wie etwa, dass längere Haare zu besseren Noten führen, müsste in einem wissenschaftlichen

Experiment überprüft werden. Tiefere Analysen sind notwendig, um den Zusammenhang zwischen den Variablen zu verstehen.

Die rechte Grafik zeigt einen negativen Zusammenhang zwischen der Nutzung von Social Media und der Zufriedenheit. Wenn die Nutzung von Social Media steigt, sinkt die Zufriedenheit. Auch hier ist Vorsicht geboten: Es könnte sein, dass Menschen, die unzufrieden sind, mehr Zeit in sozialen Medien verbringen. Es könnte aber auch sein, dass die Nutzung von Social Media selbst unglücklicher macht. Auch hier ist eine tiefere Analyse notwendig, um den Zusammenhang zwischen den Variablen zu verstehen.

In beiden Fällen kann lediglich ein Zusammenhang zwischen zwei Variablen beobachtet werden. Auf der linken Grafik erkennt man: „steigt A, steigt auch B“, bzw. „steigt B, steigt auch A“ (linke Abbildung). Auf der rechten Grafik kann beobachtet werden: „steigt A, sinkt B“, bzw. „steigt B, sinkt A“. In diesen Fällen spricht man von einer **Korrelation**. Eine Korrelation beschreibt eine statistische Beziehung zwischen zwei Variablen. Eine Korrelation bedeutet nicht zwangsläufig, dass eine Variable die andere beeinflusst. Es kann auch sein, dass beide Variablen von einer dritten Variable beeinflusst werden oder dass die Korrelation rein zufällig ist.

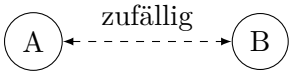
- **Positive Korrelation** bedeutet, dass wenn eine Variable steigt, auch die andere Variable steigt.
- **Negative Korrelation** bedeutet, dass wenn eine Variable steigt, die andere Variable sinkt.
- **Keine Korrelation** bedeutet, dass es keinen linearen Zusammenhang zwischen den beiden Variablen gibt.

In diesem Kapitel werden wir uns mit den Begriffen Korrelation und Kausalität beschäftigen.

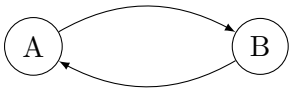
1.1 Kausalität

Definition 1.1 (Kausalität)**:**
ef:kausalitaet **Kausalität** beschreibt eine Ursache-Wirkung-Beziehung zwischen zwei Variablen, welche wir im Allgemeinen A und B nennen.
Wenn eine Variable A (Ursache) eine andere Variable B (Wirkung) beeinflusst, spricht man in diesem Fall von Kausalität. Um zu beurteilen, ob eine Kausalität vorliegt, ist es wichtig, den Zusammenhang zwischen den Variablen zu analysieren und zu verstehen. Kausalität kann häufig nur durch wissenschaftliche Experimente oder statistische Analysen nachgewiesen werden. Kausalität ist ein sehr komplexes Thema und es gibt viele verschiedene Arten von Kausalität.
Der Begriff *Kausalität* hat seinen Ursprung im Lateinischen. Das Wort entstammt dem lateinischen Begriff *causa*, was so viel wie Ursache bedeutet.
In der Regel wird Kausalität in fünf verschiedene Kategorien unterteilt:

1. **Koinzidenz** („nicht-Kausalität“) beschreibt eine zufällige Korrelation zwischen zwei Variablen, die keine kausale Beziehung haben.

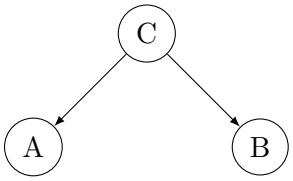
   

   Beispiel: *Der Pro-Kopf-Verbrauch von Mozzarella-Käse sowie die verliehenen Bauingenieurs-Doktortitel in den USA sind zwischen den Jahren 2000 bis 2009 beide angestiegen.*
2. **Zyklische Beeinflussung** beschreibt eine Situation, in der zwei Variablen sich gegenseitig beeinflussen (gegenseitige Verstärkung im positiven oder negativen Sinn).



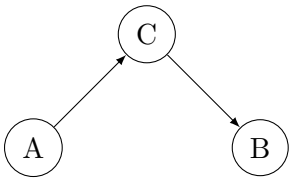
Beispiel: *Schlafmangel führt zu mehr Stress, was wiederum zu mehr Schlafmangel führt.*

3. **Gemeinsamer Grund** beschreibt eine Situation, in der eine dritte Variable C sowohl A als auch B beeinflusst.



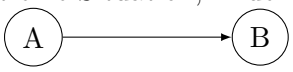
Beispiel: *Zwischen November und Januar steigen sowohl der Verkauf von Winterreifen als auch die Häufigkeit von Erkältungen an (Grund: Jahreszeit).*

4. **Indirekte Kausalität** beschreibt eine Situation, in der A B beeinflusst, aber nicht direkt, sondern via eine dritte Variable C.



Beispiel: *Sport führt zu besserem Schlaf, was wiederum zu besseren Prüfungsergebnissen führt.*

5. **Direkte Kausalität** beschreibt eine Situation, in der A B direkt beeinflusst.



Beispiel: *Mehr Video-Streams führen automatisch zu mehr Einnahmen für die Youtube-rin.*

Aufgabe 1.1

In einer Klasse wurde ein Programmier-Test durchgeführt. Die SchülerInnen mussten 10 Fragen beantworten. Für jede richtige Antwort gab es 1 Punkt, für jede falsche Antwort gab es 0 Punkte. Die SchülerInnen wurden anschliessend in eine Rangliste eingeteilt, wobei die Person mit den meisten Punkten den ersten Platz belegte. Die Tabelle **??** zeigt die Ergebnisse des Tests. Vervollständigen Sie die fehlenden Daten. Um welche Art von Zusammenhang (s. **??**) handelt es sich?

Nr.	Name	Anzahl Punkte	Rang
1	Anna Keller	2.5	5
2	Jonas Meier	4.0	?
3	Lara Schmid	1.0	7
4	Elias Huber	3.5	3
5	Mia Müller	5.0	1
6	Noah Fischer	2.0	?
7	Sofia Weber	3.0	4

Tabelle 1.1: Rangliste des Programmier-Tests
ab:rangliste

✔ Lösungsvorschlag zu Aufgabe 1.1

Der Rang lässt sich direkt aus der Anzahl Punkte ableiten: Jonas belegt den 2. Rang und Noah den 6. Rang. Es handelt sich um eine **direkte Kausalität**, da die Rangliste direkt von der Anzahl Punkte abhängt. Wenn jemand mehr Punkte hat, hat er / sie auch einen höheren Rang.

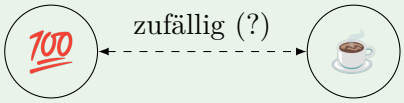
Aufgabe 1.2

Bestimmen Sie für folgende Aussagen, welche Art von Kausalität vorliegt:

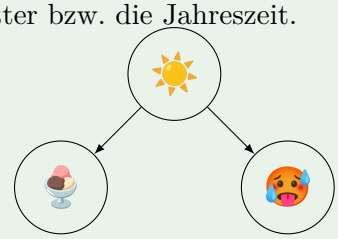
1. Die Durchschnittsnote in der Klasse steigt über das Schuljahr hinweg. Im gleichen Zeitraum haben die Lehrpersonen ihren Kaffeekonsum gesteigert.
2. In den Sommermonaten nehmen sowohl die Anzahl Sonnenbrände wie auch der Verkauf von Glacé zu.
3. In den letzten zehn Jahren hat der Konsum von Avocados in der Schweiz zugenommen. Im gleichen Jahrzehnt ist die Zahl bestandener Maturaprüfungen gestiegen.
4. Eine Studie hat gezeigt, dass HunderbesitzerInnen weniger Risiko haben, übergewichtig zu werden.
5. Eine Untersuchung von 1000 SchülerInnnen hat gezeigt, dass SchülerInnen, welche sehr gestresst sind, schlechtere Noten erzielen.

✔ Lösungsvorschlag zu Aufgabe 1.2

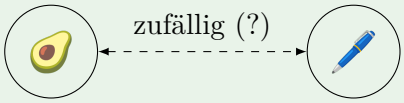
- **Koinzidenz**: Keine kausale Beeinflussung. Die Korrelation ist zufällig, oder höchstens durch eine Drittvariable erklärbar (z.B. die Lehrpersonen investieren mehr Zeit in den Unterricht → mehr Stress → mehr Kaffee).



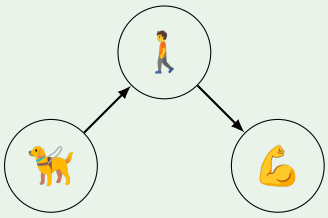
- **Gemeinsamer Grund**: Beide Grössen werden durch eine gemeinsame Drittvariable beeinflusst: das Wetter bzw. die Jahreszeit.



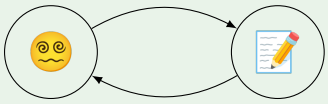
- **Koinzidenz**: Keine echte Beeinflussung. Es handelt sich um eine Scheinkorrelation ohne sinnvollen Zusammenhang. Der gleichzeitige Trend ist rein zufällig, es sei denn, die Forschung zeige auf, dass Avocados eine positive Wirkung auf die kognitiven Fähigkeiten haben.



- **Indirekte Kausalität**: HundebesitzerInnen bewegen sich mehr, was zu weniger Übergewicht führt. Es handelt sich um eine indirekte Kausalität.



- **Zyklische Beeinflussung**: Stress führt zu schlechteren Noten, was wiederum zu mehr Stress führt. Es handelt sich um eine zyklische Beeinflussung.



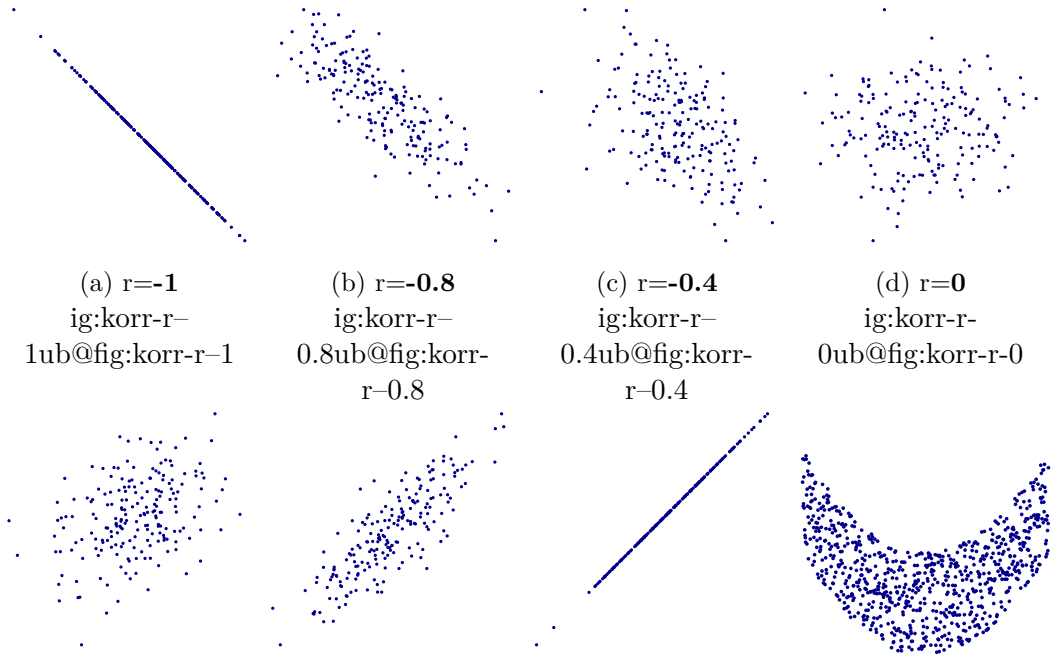
1.2 Korrelation

Definition 1.2 (Korrelation)**:**
Korrelation beschreibt eine statistische Beziehung zwischen zwei Variablen. Das Wort Korrelation kommt von den lateinischen Wörtern *co(n)*-=„mit“ und *relatio* = Beziehung, was soviel wie eine gegenseitige Beziehung oder auch Wechselbeziehung bedeutet (🔄). Im Kern bedeutet dies lediglich eine statistische Beziehung zwischen zwei Variablen (Beobachtungen) A und B.
Eine Korrelation bedeutet nicht zwangsläufig, dass eine Variable die andere beeinflusst (s.

??). Es kann auch sein, dass beide Variablen von einer dritten Variable beeinflusst werden oder dass die Korrelation rein zufällig ist.

- **Positive Korrelation** bedeutet, dass wenn eine Variable steigt, auch die andere Variable steigt.
- **Negative Korrelation** bedeutet, dass wenn eine Variable steigt, die andere Variable sinkt.
- **Keine Korrelation** bedeutet, dass es keinen linearen Zusammenhang zwischen den beiden Variablen gibt.

Der Korrelationskoeffizient wird typischerweise als Zahl r angegeben, welche sich zwischen -1 (perfekte negative Korrelation) und 1 (perfekte positive Korrelation) bewegt (s. **??** und **??**). Ein Wert von 0 bedeutet, dass es keinen linearen Zusammenhang zwischen den beiden Variablen gibt (s. **??**). Dies schliesst jedoch nicht aus, dass es einen anderen, nicht-linearen Zusammenhang gibt (s. **??**).



(a) r=**-1** ig:korr-r–1ub@fig:korr-r–1

(b) r=**-0.8** ig:korr-r–0.8ub@fig:korr-r–0.8

(c) r=**-0.4** ig:korr-r–0.4ub@fig:korr-r–0.4

(d) r=**0** ig:korr-r-0ub@fig:korr-r-0

(e) r=**0.4** ig:korr-r-0.4ub@fig:korr-r-0.4

(f) r=**0.8** ig:korr-r-0.8ub@fig:korr-r-0.8

(g) r=**1** ig:korr-r-1ub@fig:korr-r-1

(h) r≃**0** ig:korr-r-oth0ub@fig:korr-r-oth0

Abbildung 1.2: Beispiele für den Korrelationskoeffizienten r
ig:korr-r-bsp

?? zeigt zwei Beispiele für Korrelationen. In der linken Grafik ist eine positive Korrelation zwischen der Haarlänge und dem Notendurchschnitt zu sehen. In der rechten Grafik ist eine negative Korrelation zwischen der Nutzung von Social Media und der Zufriedenheit zu sehen.

Um Zusammenhänge zwischen zwei Variablen nicht nur zu benennen, sondern auch quantifizieren und vergleichen zu können, wäre es hilfreich, den Zusammenhang zwischen zwei Datenreihen in einer einzigen Zahl r ausdrücken zu können. Hilfreich wäre es zudem, falls diese Zahl positiv wäre bei positiven Korrelationen und negativ bei negativer Korrelation. Damit könnten Fragen wie die Folgende beantwortet werden:

„Welcher Zusammenhang ist stärker – Schlafdauer vs. Konzentration (r = 0.6) oder Handynutzung

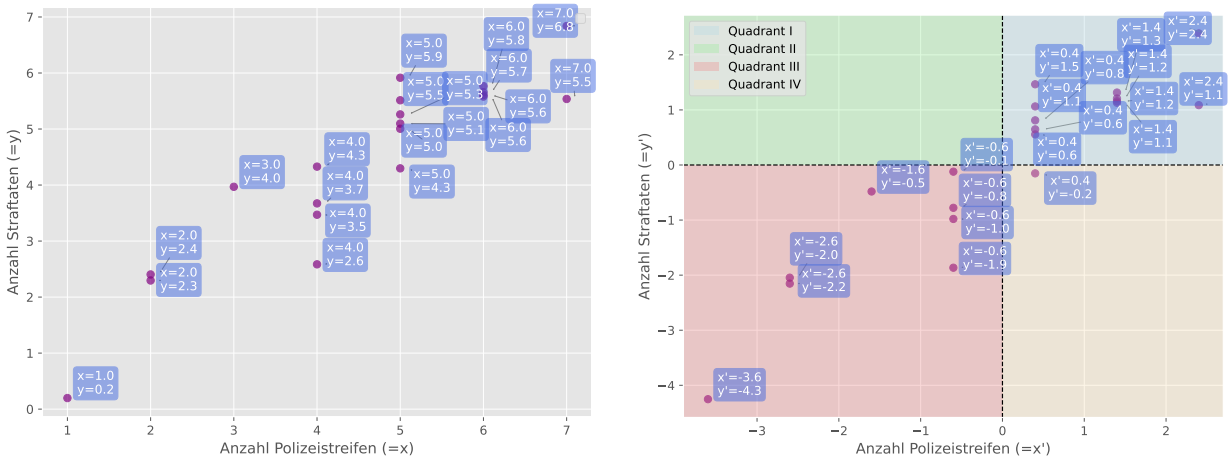
vs. Noten (r = –0.4)?“

In Psychologie, Wirtschaft, Medizin oder Sport begegnet man Korrelationen oft: z.B. “Bewegung korreliert mit geringerer Krankheitsanfälligkeit”. Wer r versteht, kann solche Aussagen kritisch beurteilen und Studienergebnisse kritisch interpretieren.

Die Stärke der Korrelation zwischen zwei Variablen kann durch eine mathematische Formel berechnet werden. Eine Möglichkeit, dies zu tun, ist die Berechnung des Korrelationskoeffizienten r.

Der erste Schritt zur Berechnung des Korrelationskoeffizienten ist die Berechnung der Mittelwerte aller Variablen, $\overline{x}$ und $\overline{y}$ (s. **??**). Der Mittelwert ist der arithmetische Durchschnittswert aller Werte einer Variablen.

In einem zweiten Schritt berechnen wir die Abweichung der ursprünglichen Werte von ihrem Mittelwert (s. **??**). Die Abweichung eines Wertes x_i von seinem Mittelwert $\overline{x}$ wird als x'_i bezeichnet.



(a) Ursprüngliche Daten ig:Korrelationen-Ursprungub@fig:Korrelationen-Ursprung

(b) Daten als Abweichung vom Mittelwert ig:Korrelationen-Abweichung-Ursprungub@fig:Korrelationen-Abweichung-Ursprung

Abbildung 1.3: Korrelations-Beispiel mit Abweichung vom Mittelwert ig:Korrelationen-Abweichung

Wie man in **??** erkennen kann, sieht die Abbildung der Daten als Abweichung vom Mittelwert fast gleich wie die ursprünglichen Werte selber aus, ausser dass die Werte jetzt um 0 zentriert sind. Somit kann man folgenden „Trick“ anwenden, um den Korrelationskoeffizienten zu berechnen: Wir multiplizieren die Abweichungen der beiden Variablen miteinander und addieren die Produkte auf. Das Ergebnis ist eine Zahl, die den Zusammenhang zwischen den beiden Variablen beschreibt. Wenn die Daten im roten oder blauen Bereich liegen, ist das Produkt $x'_i \cdot y'_i$ positiv. Wenn die Daten im grünen oder gelben Bereich liegen, ist das Produkt negativ.

Wir berechnen nun also die Zahl Z:

$$Z = x'_1 \cdot y'_1 + x'_2 \cdot y'_2 + \cdots + x'_n \cdot y'_n$$

Aufgabe 1.3

Verwenden Sie die Excel-Datei auf Moodle und berechnen Sie die Zahl Z für die beiden Datenserien. Welche Korrelation liegt vor? Ist es eine positive oder negative Korrelation?

Lösungsvorschlag zu Aufgabe 1.3

- Die Korrelation zwischen der Berufserfahrung (Jahre) und der Einladungswahrscheinlichkeit beträgt $Z = 44.66$. Es handelt sich um eine positive Korrelation.
- Die Korrelation zwischen der Nutzung von Social Media und der Zufriedenheit beträgt $Z = -30.13$. Es handelt sich um eine negative Korrelation.

Aufgabe 1.4

Verwandeln Sie in Aufgabe Aufgabe 1.3 die Anzahl Jahre Berufserfahrung in eine neue Variable, die die Anzahl Jahre Berufserfahrung in Monaten angibt. Berechnen Sie die Zahl Z für die beiden Datenserien. Wie verändert sich die Korrelation?

Lösungsvorschlag zu Aufgabe 1.4

Die Korrelation zwischen der Berufserfahrung (Monate) und der Einladungswahrscheinlichkeit beträgt nun neu $Z = 18.78$. Der Korrelationswert hat sich also nur aufgrund einer andere Grösseneinheit verändert. Dies wirft die Frage auf, wie man einen Korrelationskoeffizienten erschaffen könnte, der nicht von der Grösseneinheit abhängt.

Wie wir in Aufgabe 1.4 gesehen haben, ist die Zahl Z von der Grösseneinheit abhängig. Das bedeutet, dass wir die Stärke der Korrelation nicht direkt vergleichen können, wenn die Variablen unterschiedliche Grösseneinheiten haben. Um dies zu lösen, verwenden wir den **Korrelationskoeffizienten** r, welcher die Zahl Z zwischen -1 und 1 normiert:

$$\begin{aligned} r &= \frac{Z}{\sqrt{\left((x'_1)^2 + (x'_2)^2 + \cdots + (x'_n)^2\right) \cdot \left((y'_1)^2 + (y'_2)^2 + \cdots + (y'_n)^2\right)}} \\ &= \frac{Z}{\sqrt{\left(\sum_{i=1}^{n} (x'_i)^2\right) \cdot \left(\sum_{i=1}^{n} (y'_i)^2\right)}} \end{aligned} \tag{1.1}$$

Der Korrelationskoeffizient r ist eine standardisierte Version der Zahl Z. Der Korrelationskoeffizient r liegt immer zwischen -1 und 1. Ein Wert von 1 bedeutet, dass die beiden Variablen in einem perfekten positiven linearen Zusammenhang stehen. Ein Wert von -1 bedeutet, dass die beiden Variablen in einem perfekten negativen linearen Zusammenhang stehen. Ein Wert von 0 bedeutet, dass es keinen linearen Zusammenhang zwischen den beiden Variablen gibt. Dies schliesst jedoch nicht aus, dass es einen anderen, nicht-linearen Zusammenhang gibt.

Aufgabe 1.5

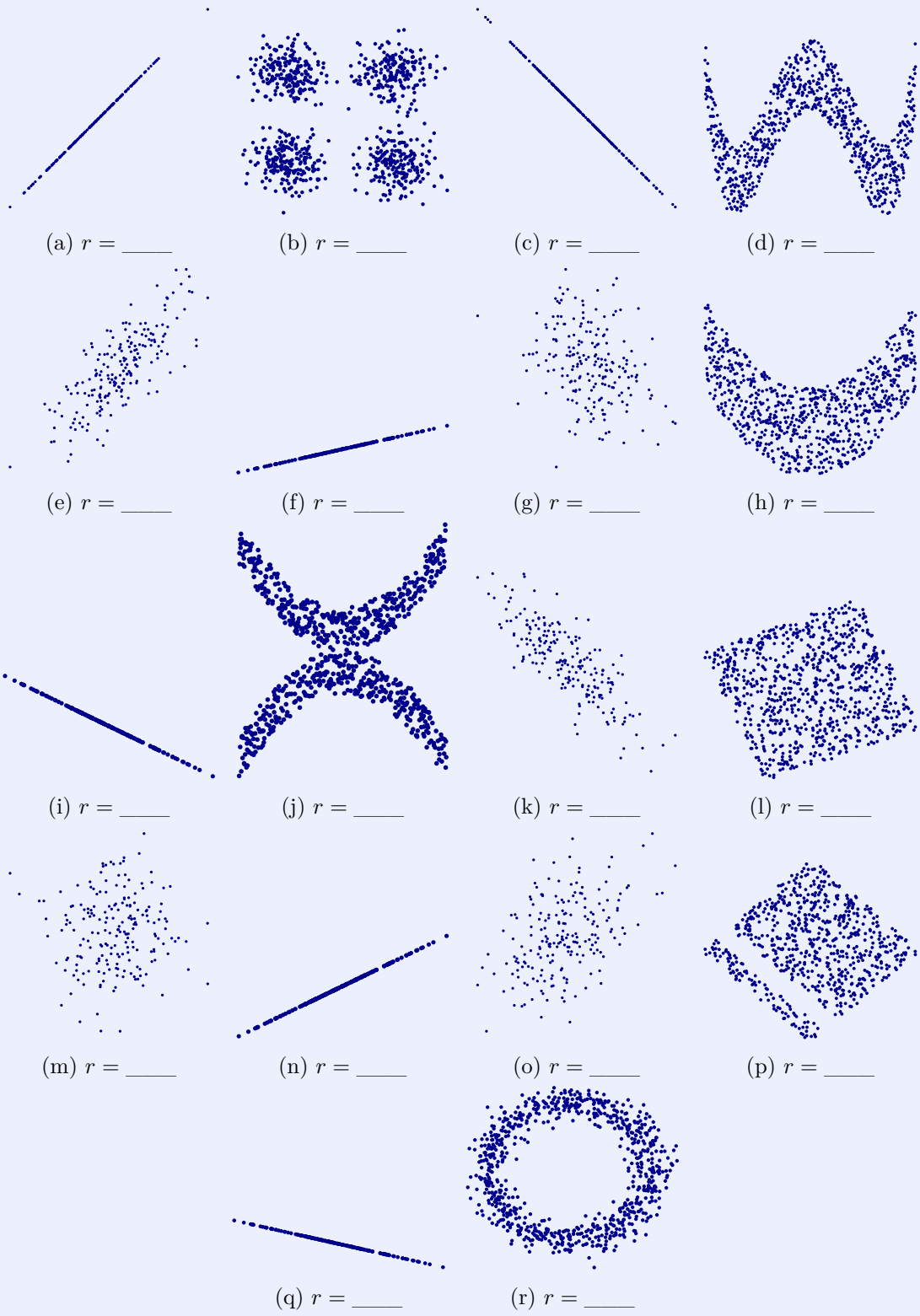
Berechnen Sie den Korrelationskoeffizienten r für die beiden Datenserien in Aufgabe 1.3. Wie stark ist die jeweilige Korrelation?

Lösungsvorschlag zu Aufgabe 1.5

- Die Korrelation zwischen der Berufserfahrung (Jahre) und der Einladungswahrscheinlichkeit beträgt $r = 0.96$. Es handelt sich um eine sehr starke positive Korrelation.
- Die Korrelation zwischen der Nutzung von Social Media und der Zufriedenheit beträgt $r = -0.83$. Es handelt sich um eine starke negative Korrelation.

Aufgabe 1.6

Ordnen Sie die folgenden Bilder steigend nach dem Korrelationskoeffizienten r ein. Welches Bild hat den höchsten Korrelationskoeffizienten? Welches Bild hat den tiefsten Korrelationskoeffizienten?



(a) $r =$ _____ (b) $r =$ _____ (c) $r =$ _____ (d) $r =$ _____

(e) $r =$ _____ (f) $r =$ _____ (g) $r =$ _____ (h) $r =$ _____

(i) $r =$ _____ (j) $r =$ _____ (k) $r =$ _____ (l) $r =$ _____

(m) $r =$ _____ (n) $r =$ _____ (o) $r =$ _____ (p) $r =$ _____

(q) $r =$ _____ (r) $r =$ _____

Aufgabe 1.7

Luca und Mira bereiten sich auf eine Prüfung vor. Sie erfassen während der Vorbereitung drei Grössen:

- **B: B**eginn der Lernphase (z. B. 8.5 = 8:30 Uhr)
- **D: D**auer der Lerneinheit in Stunden
- **A: A**nzahl bearbeiteter Seiten pro Tag

Die folgende Tabelle zeigt die Korrelationen zwischen diesen drei Variablen:

	B und D	**B und A**	**D und A**
Luca	−0.69	−0.04	0.25
Mira	0.64	0.54	0.89

1. Beantworten Sie folgenden Fragen mit „Luca“, „Mira“, „beide“, oder „keine“.:
   (a) Wer lernt länger, wenn er oder sie ausgeschlafen ist? *(B kleiner = früherer Start)*
   (b) Bei wem hat der Zeitpunkt des Lernstarts keinen Einfluss auf die Anzahl bearbeiteter Seiten?
   (c) Bei wem führt eine längere Dauer der Lerneinheit zu einer höheren Anzahl bearbeiteter Seiten?
2. **Wer braucht insgesamt weniger Zeit, um den gesamten Stoff durchzuarbeiten?**

Lösungsvorschlag zu Aufgabe 1.7

1. Erste Fragen
   (a) Bei B und D gibt es bei **Mira** eine positive Korrelation: Wenn sie länger schläft (später aufsteht), lernt sie länger.
   (b) Bei B und A gibt es bei **Luca** eine Korrelation von fast 0: Der Zeitpunkt des Lernstarts hat keinen Einfluss auf die Anzahl bearbeiteter Seiten.
   (c) Bei **beiden** gibt es eine positive Korrelation: Wenn die Dauer der Lerneinheit steigt, steigt auch die Anzahl bearbeiteter Seiten. Bei Mira ist die Korrelation jedoch viel stärker (0.89) als bei Luca (0.25).
2. Das kann aufgrund der gegebenen Information nicht mit Sicherheit gesagt werden. Man bräuchte dazu genauere Kenntnisse über die Daten.

1.3 Lernziele

- ☐ Ich kann den Unterschied zwischen Korrelation und Kausalität erklären.
- ☐ Ich kann begründete Einschätzungen über die Kausationsart zwischen zwei Variablen abgeben, indem ich folgende Begriffe verwende:
  - Koinzidenz
  - Zyklische Beeinflussung
  - Gemeinsamer Grund
  - Indirekte Kausalität
  - Direkte Kausalität
- ☐ Ich wende bei der Interpretation von korrelierten Daten Vorsicht an, bevor ich Aussagen über Kausalität treffe.
- ☐ Ich kann den Korrelationskoeffizienten Z berechnen und interpretieren.
- ☐ Ich kann unterschiedliche Abbildungen von Daten betreffend ihrem geschätzten Z-Wert vergleichen.
- ☐ Ich kann den Korrelationskoeffizienten r berechnen und interpretieren.

Kapitel 2

Lineare Regression

Eine häufige Aufgabe im Bereich der künstlichen Intelligenz besteht darin, für unbekannte Datenpunkte Vorhersagen zu treffen. In diesem Kapitel beschäftigen wir uns mit der linearen Regression, einer der grundlegendsten und am häufigsten verwendeten Methoden zur Vorhersage von Werten.

2.1 Grundlagen der linearen Regression

Definition 2.1 (Lineare Regression)**:**
Lineare Regression ist ein Verfahren zur Vorhersage einer abhängigen Variablen auf Basis einer unabhängigen Variablen. Dabei wird eine Gerade (lineare Funktion) durch die Datenpunkte gelegt, die den besten Fit (die beste „Passung“) zu den Datenpunkten hat. Diese Gerade wird als Regressionsgerade bezeichnet.
Die Regressionsgerade hat die Form: $y = mx + q$, wobei:

- y die abhängige Variable ist (die Variable, die wir vorhersagen wollen)
- x die unabhängige Variable ist (die Variable, die wir zur Vorhersage verwenden)
- m die Steigung der Geraden ist
- q der y-Achsenabschnitt ist (der Wert von y, wenn $x = 0$)

Der Begriff *Regression* wurde erstmals vom britischen Statistiker Francis Galton verwendet, der die „Rückkehr zum Mittelwert“ (engl. „regression toward the mean“) beschrieb. In der modernen Statistik bezieht sich der Begriff auf verschiedene Methoden zur Modellierung von Beziehungen zwischen Variablen.

?? zeigt ein Beispiel für ein Streudiagramm, das den Zusammenhang zwischen Schlafdauer und Note darstellt. Wir können beobachten, dass es einen positiven Zusammenhang zwischen diesen beiden Variablen gibt: Je mehr Schlaf, desto besser die Note. Aber wie können wir diesen Zusammenhang nutzen, um Vorhersagen zu treffen?

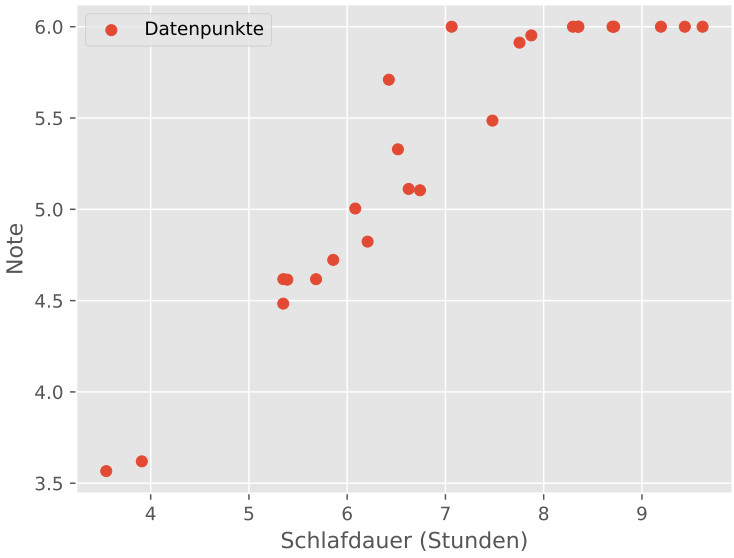


Abbildung 2.1: Schlafdauer und Note
ig:schlafdauer$_v s_n ote$

Beispielsweise könnten wir versuchen, folgende Frage zu beantworten: **Welche Note werde ich erreichen, wenn ich im Schnitt 8 Stunden pro Nacht schlafe?**

Bevor wir mit der linearen Regression beginnen, sollten jedoch noch die Begriffe **unabhängige** und **abhängige Variable** geklärt werden.

Definition 2.2 (Unabhängige und abhängige Variable)**:**

- Die **unabhängige Variable** (auch Prädiktor- oder Eingabevariable genannt) ist die Variable, die wir kontrollieren oder messen, um eine Vorhersage zu treffen. In unserem Beispiel ist die Schlafdauer die unabhängige Variable.
- Die **abhängige Variable** (auch Zielvariable oder Ausgabevariable genannt) ist die Variable, die wir vorhersagen wollen. In unserem Beispiel ist die Note die abhängige Variable.

In Streudiagrammen wird üblicherweise die unabhängige Variable auf der x-Achse und die abhängige Variable auf der y-Achse dargestellt.

2.2 Einfache lineare Regression

2.2.1 Eine erste Näherung

Um einen ersten Eindruck zu bekommen, wie eine Regressionsgerade funktioniert, betrachten wir einen einfachen Ansatz: Wir verbinden den Datenpunkt mit der kleinsten x-Koordinate mit dem Datenpunkt mit der grössten x-Koordinate.

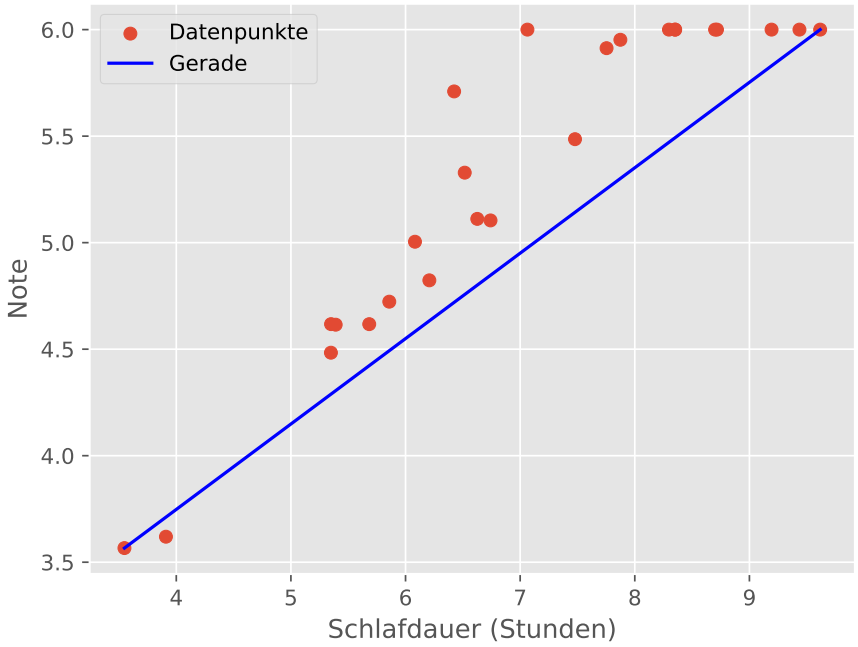


Abbildung 2.2: Einfache Regressionsgerade durch Verbinden des ersten und letzten Punktes ig:simple$_r egression_l ine$

Um diese Gerade zu bestimmen, benötigen wir ihre Steigung und ihren y-Achsenabschnitt:

- Die Steigung m berechnet sich aus: $m = \frac{y_2 - y_1}{x_2 - x_1}$, wobei (x_1, y_1) und (x_2, y_2) die beiden miteinander verbundenen Punkte sind.
- Der y-Achsenabschnitt q berechnet sich aus: $q = y_1 - m \cdot x_1$

Mit dieser Geraden können wir nun Vorhersagen treffen: Für einen gegebenen x-Wert (z.B. 8 Stunden Schlaf) berechnen wir den entsprechenden y-Wert (die vorhergesagte Note) mit der Formel $y = m \cdot x + q$.

Aufgabe 2.1

Gegeben seien die Punkte $(3.54, 3.57)$ und $(9.62, 6.0)$, die die Schlafdauer in Stunden und den Notenschnitt darstellen.

1. Berechnen Sie die Steigung m und den y-Achsenabschnitt q der Geraden durch diese Punkte.
2. Welchen Notenschnitt würde man bei 7 Stunden Schlaf erwarten?

Lösungsvorschlag zu Aufgabe 2.1

1. Steigung: $m = \frac{6.0 - 3.57}{9.62 - 3.54} = \frac{2.43}{6.08} \approx 0.40$
   y-Achsenabschnitt: $q = 3.57 - 0.40 \cdot 3.54 = 3.57 - 1.416 \approx 2.15$
   Die Geradengleichung lautet also: $y = 0.40x + 2.15$
2. Bei 7 Stunden Schlaf: $y = 0.40 \cdot 7 + 2.15 = 2.80 + 2.15 = 4.95$
   Man würde einen Notenschnitt von 4.95 erwarten.

Diese Methode ist einfach, hat aber einen grossen Nachteil: Sie berücksichtigt nur zwei Datenpunkte und ignoriert alle anderen Datenpunkte. Dadurch kann sie sehr anfällig für Ausreisser sein.

2.2.2 Methode der kleinsten Quadrate

Eine bessere Methode zur Bestimmung der Regressionsgerade ist die **Methode der kleinsten Quadrate** (engl. *Ordinary Least Squares*, OLS). Diese Methode wählt die Gerade so, dass die Summe der quadrierten Abweichungen zwischen den tatsächlichen y-Werten und den vorhergesagten y-Werten minimal ist.

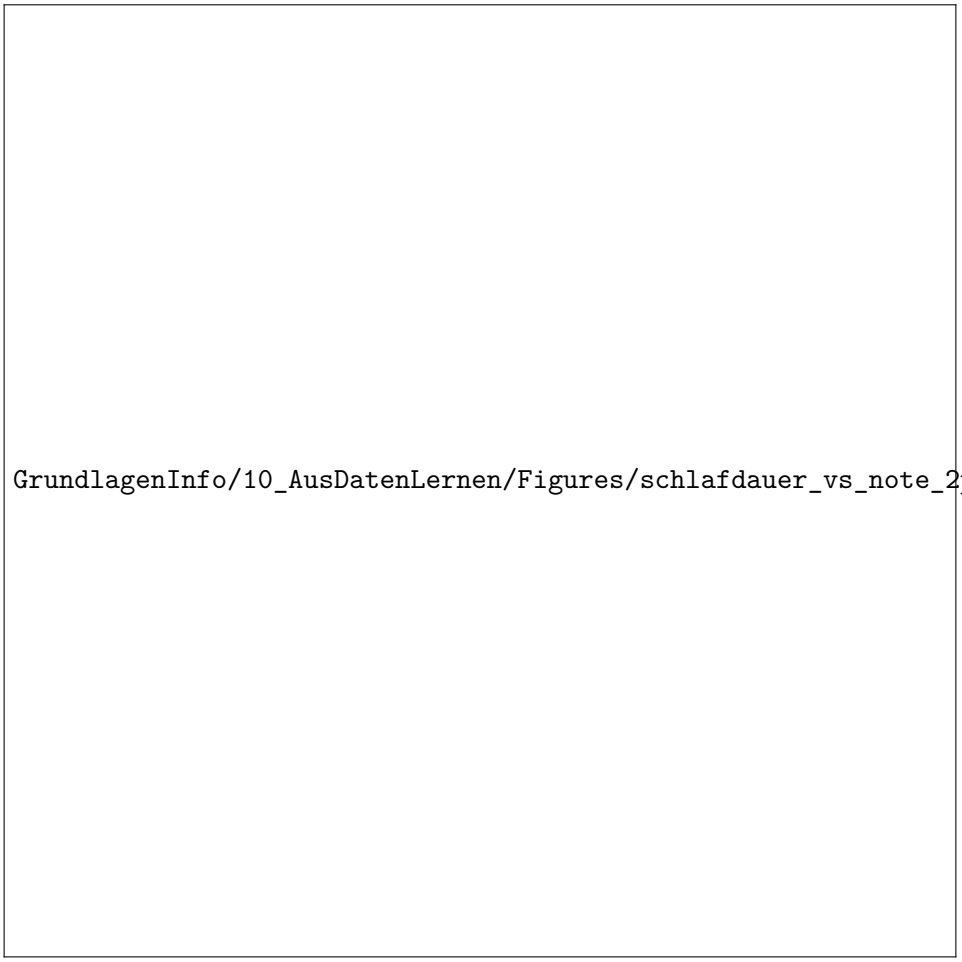


Abbildung 2.3: Darstellung der Residuen (Abweichungen) bei der linearen Regression
ig:residuals

Diese Abweichungen werden als **Residuen** bezeichnet. Das Ziel der Methode der kleinsten Quadrate ist es, die Summe der quadrierten Residuen zu minimieren:

$$S = \sum_{i=1}^{n} (y_i - (mx_i + q))^2$$

wobei (x_i, y_i) die Datenpunkte sind und n die Gesamtzahl der Datenpunkte.

Um die Werte für m und q zu finden, die S minimieren, setzen wir die partiellen Ableitungen von S nach m und q gleich Null und lösen das resultierende Gleichungssystem (s. Kapitel A). Dies führt zu den folgenden Formeln:

$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$q = \frac{\sum y_i - m \sum x_i}{n}$$

Aufgabe (Challenge) 2.2

Betrachten Sie die folgenden Datenpunkte, die die Schlafdauer (in Stunden) und den Notenschnitt darstellen:

Schlafdauer (x)	5	6	7	8	9
Notenschnitt (y)	3.8	4.2	4.8	5.1	5.6

1. Berechnen Sie die Steigung m und den y-Achsenabschnitt q mit der Methode der kleinsten Quadrate.
2. Welchen Notenschnitt würde man bei 7.5 Stunden Schlaf erwarten?

Lösungsvorschlag zu Aufgabe 2.2

1. Um die Werte für m und q zu berechnen, bestimmen wir zuerst die Summen:
   $n = 5$
   $\sum x_i = 5 + 6 + 7 + 8 + 9 = 35$
   $\sum y_i = 3.8 + 4.2 + 4.8 + 5.1 + 5.6 = 23.5$
   $\sum x_i y_i = 5 \cdot 3.8 + 6 \cdot 4.2 + 7 \cdot 4.8 + 8 \cdot 5.1 + 9 \cdot 5.6 = 19 + 25.2 + 33.6 + 40.8 + 50.4 = 169$
   $\sum x_i^2 = 5^2 + 6^2 + 7^2 + 8^2 + 9^2 = 25 + 36 + 49 + 64 + 81 = 255$
   Nun können wir m berechnen: $m = \frac{5 \cdot 169 - 35 \cdot 23.5}{5 \cdot 255 - 35^2} = \frac{845 - 822.5}{1275 - 1225} = \frac{22.5}{50} = 0.45$
   Und q: $q = \frac{23.5 - 0.45 \cdot 35}{5} = \frac{23.5 - 15.75}{5} = \frac{7.75}{5} = 1.55$
   Die Regressionsgleichung lautet also: $y = 0.45x + 1.55$
2. Bei 7.5 Stunden Schlaf: $y = 0.45 \cdot 7.5 + 1.55 = 3.375 + 1.55 = 4.925$
   Man würde einen Notenschnitt von 4.93 erwarten.

2.3 Implementierung mit Python

In der Praxis wird die lineare Regression oft mit statistischen Softwarepaketen oder Programmiersprachen wie Python implementiert. In Python gibt es mehrere Bibliotheken, die lineare Regressionen unterstützen, darunter `statsmodels` und `scikit-learn`.

Hier ist ein Beispiel für die Implementierung einer linearen Regression mit `statsmodels`:

```
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt

# Daten erstellen
data = {
    'Schlafdauer': [5, 6, 7, 8, 9],
    'Notenschnitt': [3.8, 4.2, 4.8, 5.1, 5.6]
}
df = pd.DataFrame(data)

# Lineare Regression durchführen
model = smf.ols(formula="Notenschnitt ~ Schlafdauer", data=df).fit()
```

```
# Ergebnisse ausgeben
print(model.params)
```

Die Ergebnisse der Regression enthalten die Steigung (`Schlafdauer`) und den y-Achsenabschnitt (`Intercept`):

```
Intercept      1.55
Schlafdauer    0.45
dtype: float64
```

2.4 Qualität einer Regression

Wie gut passt eine Regressionsgerade zu den Daten? Um diese Frage zu beantworten, betrachten wir insbesondere die Residuen und die Summe der quadrierten Residuen

Wie bereits erwähnt, sind Residuen die Abweichungen zwischen den tatsächlichen y-Werten und den durch die Regressionsgerade vorhergesagten y-Werten. Die Summe der quadrierten Residuen (auch als Residuenquadratsumme oder RSS bezeichnet) ist ein Mass für die Gesamtabweichung:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

wobei $\hat{y}_i = mx_i + q$ der vorhergesagte Wert für den i-ten Datenpunkt ist.

Je kleiner die RSS, desto besser passt die Regressionsgerade zu den Daten.

2.5 Anwendungen und Grenzen der linearen Regression

2.5.1 Anwendungen

Die lineare Regression hat vielfältige Anwendungen in verschiedenen Bereichen, beispielsweise:

- **Wirtschaft**: Vorhersage von Verkaufszahlen, Aktienkursen oder wirtschaftlichen Indikatoren
- **Medizin**: Untersuchung des Zusammenhangs zwischen Risikofaktoren und Krankheiten
- **Psychologie**: Analyse von Beziehungen zwischen psychologischen Variablen
- **Maschinelles Lernen**: Grundlage für komplexere Vorhersagemodelle

2.6 Lernziele

- ☐ Ich kann erklären, was lineare Regression ist und wofür sie verwendet wird.
- ☐ Ich kann zwischen unabhängigen und abhängigen Variablen unterscheiden.
- ☐ Ich kann die Regressionsgerade für einen Datensatz manuell berechnen.
- ☐ Ich kann die Qualität einer Regression beurteilen.
- ☐ Ich kann eine lineare Regression mit Python implementieren.
- ☐ Ich kenne die Anwendungen und Grenzen der linearen Regression.

Kapitel 3

Zustände Vorhersagen mit Markov-Ketten

3.1 Einleitung

Viele Vorgänge in unserem Alltag lassen sich als Abfolge von Zuständen beschreiben: Das Wetter ändert sich von Tag zu Tag, Menschen wechseln zwischen verschiedenen Apps auf ihrem Smartphone oder ein Zug ist mal pünktlich, mal verspätet. Dabei interessiert uns oft, wie wahrscheinlich der nächste Zustand ist. Hier setzen **Markov-Ketten** an. Diese können genutzt werden, um Zustände (z.B. Wetterzustände, App-Nutzung) in einem zeitlichen Verlauf zu modellieren und Vorhersagen zu treffen.

Eine Markov-Kette ist ein mathematisches Modell zur Beschreibung solcher Prozesse. Ihr zentrales Merkmal ist die sogenannte *Markov-Eigenschaft*: Der nächste Zustand hängt nur vom aktuellen Zustand ab – nicht davon, wie man in diesen Zustand gekommen ist. Das macht Markov-Ketten zu einem besonders einfachen und sehr verbreiteten Werkzeug.

Das Prinzip von Markov-Ketten lässt sich gut am Beispiel des Wetters illustrieren. Wenn es heute sonnig ist, so können Sie mit einer gewissen Wahrscheinlichkeit abschätzen, wie das Wetter morgen sein wird – zum Beispiel wieder sonnig oder regnerisch. Die genaue Wahrscheinlichkeit hängt nur vom heutigen Wetter ab, nicht von den Wetterverhältnissen der letzten Tage.

Aufgabe 3.1

Entscheiden Sie für jede der folgenden Situationen, ob die Markov-Eigenschaft erfüllt ist. Begründen Sie Ihre Antwort jeweils kurz.

1. Das Wetter morgen kann gut vorhergesagt werden basierend auf dem heutigen Wetter.
2. Die Wahrscheinlichkeit, dass ein Schüler heute seine Hausaufgaben macht, hängt davon ab, ob er sie gestern und vorgestern gemacht hat.
3. Die Wahrscheinlichkeit, dass ein Programm abstürzt, hängt davon ab, wie viele andere Programme bereits abgestürzt sind.
4. Die Wahrscheinlichkeit, dass eine Person Kleider einer bestimmten Marke kauft, hängt nur davon ab, ob das zuletzt gekaufte Kleidungsstück von dieser Marke war.

✔ Lösungsvorschlag zu Aufgabe 3.1

1. **Markov-Eigenschaft erfüllt**, da das Wetter morgen nur vom heutigen Wetter abhängt.
2. **Nicht erfüllt**, da die Entscheidung vom Zustand der letzten zwei Tage abhängt.
3. **Nicht erfüllt**, da die Absturz-Wahrscheinlichkeit von der gesamten Laufzeit abhängt, nicht nur vom aktuellen Zustand.
4. **Erfüllt**, da die Kaufentscheidung nur vom zuletzt gekauften Kleidungsstück abhängt.

In diesem Kapitel werden Sie anhand lebensnaher Beispiele wie Wettermodellen, App-Nutzung oder Pendlerverhalten schrittweise verstehen, wie Markov-Ketten funktionieren, wie man sie mathematisch beschreibt und wie Sie sie nutzen können, um Prognosen zu erstellen.

Als erstes müssen wir eine geeignete Weise finden, um Übergangswahrscheinlichkeiten darzustellen. Eine erste Möglichkeit wäre, diese als **gerichteten Graphen** darzustellen.

3.2 Darstellung von Übergangswahrscheinlichkeiten als gerichteter Graph

Markov-Ketten lassen sich gut als gerichtete Graphen visualisieren, wobei die Zustände als Knoten und die Übergangswahrscheinlichkeiten als beschriftete Kanten dargestellt werden.

Beispiel 3.1 (Darstellung einer Markov-Kette als **gerichteter Graph**)**:**
xample:graph Stellen Sie sich vor, das Wetter kann nur zwei Zustände annehmen: *sonnig* oder *regnerisch*. Die Übergangswahrscheinlichkeiten sind wie folgt:

- Ist es heute sonnig, so ist die Wahrscheinlichkeit, dass es morgen wieder sonnig ist, 0.8; die Wahrscheinlichkeit für Regen beträgt 0.2.
- Ist es heute regnerisch, so ist die Wahrscheinlichkeit, dass es morgen wieder regnet, 0.6; die Wahrscheinlichkeit für Sonne beträgt 0.4.

Diese Übergangswahrscheinlichkeiten können gut als gerichteter Graph dargestellt werden:

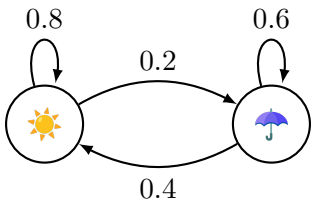


Abbildung 3.1: Visualisierung des Wetter-Beispiels als gerichteter Graph
ig:markov-weather

Dabei bezeichnen:
◯ **Knoten** = Zustände
➔ **Kanten** = Übergänge mit Wahrscheinlichkeiten

Bei dieser Darstellung wird besonders deutlich, wie die Übergänge zwischen den Zuständen erfolgen. Die Kantenbeschriftungen geben die Übergangswahrscheinlichkeiten an, wobei die ausgehenden

Kanten eines Knotens in der Summe immer 1 ergeben müssen.

Aufgabe 3.2 Visualisierung der Thailand-Markov-Kette

Ein Tourist kommt in Thailand an und möchte die Sonne und Strände auf zwei beliebten Inseln im Süden geniessen (Samui und Phangan), sowie das Festland besichtigen.

Laut einer Umfrage gilt:

- Ist der Tourist auf dem Festland, so fährt er am nächsten Tag mit Wahrscheinlichkeit 70% nach Samui, mit 20% nach Phangan und bleibt mit 10% auf dem Festland.
- Ist er auf Samui, bleibt er mit 40% dort, fährt mit 50% nach Phangan und kehrt mit 10% aufs Festland zurück.
- Ist er auf Phangan, bleibt er mit 30% dort, fährt mit 30% nach Samui und kehrt mit 40% aufs Festland zurück.

Zeichnen Sie die Markov-Kette als gerichteten Graphen. Die Zustände sind **F** (= Festland), **S** (= Samui) und **P** (= Phangan).

Lösungsvorschlag zu Aufgabe 3.2

Die Visualisierung als gerichteter Graph sieht wie folgt aus:

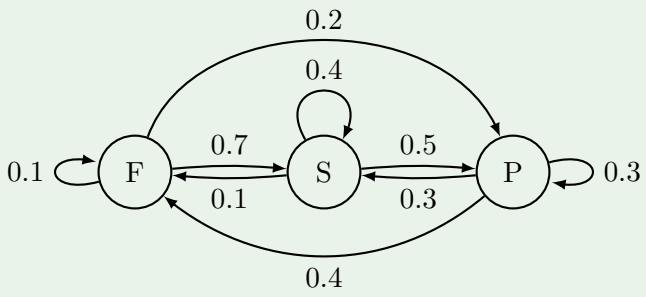
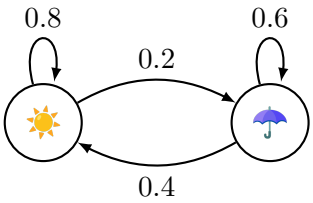


Abbildung 3.2: Visualisierung des Thailand-Beispiels
ig:markov-thailand

Wie Aufgabe 3.2 zeigt, können gerichtete Graphen schnell unübersichtlich werden. Zudem stellt sich die Frage, wie sich gerichtete Graphen so umsetzen lassen, dass ein Computer diese interpretieren kann. Aus diesen Gründen werden die Übergangswahrscheinlichkeiten von Markov-Ketten häufig als Übergangsmatrizen dargestellt.

3.3 Darstellung von Übergangswahrscheinlichkeiten als Matrix

Beispiel 3.2 (Übergangsmatrix)**:**
xample:matrix
Die Übergangswahrscheinlichkeiten aus **??** lassen sich in einer sogenannten *Übergangsmatrix* darstellen (mit Emojis, ☀️ = sonnig, ☂️ = regnerisch):



Gerichteter Graph

⇒

Von \ Nach	☀	☂
☀	0.8	0.2
☂	0.4	0.6

Übergangs-Tabelle

Die Zeilen entsprechen dem aktuellen Zustand (☀, ☂), die Spalten dem nächsten (in diesem Fall morgigen) Zustand. Häufig wird die Übergangs-Tabelle als **Übergangsmatrix** P ohne Zeilen- und Spaltenbeschriftung dargestellt:

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$$

Aufgabe 3.3 Übergangsmatrix für thailändischen Tourist

Erstellen Sie die **Übergangsmatrix** für die Thailand-Aufgabe (Aufgabe 3.2). Verwenden Sie folgende Vorlage:

Von \ Nach	F	S	P
F			
S			
P			

Lösungsvorschlag zu Aufgabe 3.3

Von \ Nach	F	S	P
F	0.1	0.7	0.2
S	0.1	0.4	0.5
P	0.4	0.3	0.3

Übergangs-Tabelle für F (Festland), S (Samui), P (Phangan)

Als Matrix und ohne Labels geschrieben:

$$P = \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.4 & 0.3 & 0.3 \end{pmatrix}$$

3.4 Vorhersage zukünftiger Zustände

Beispiel 3.3 (Vorhersage zukünftiger Zustände, basierend auf **??**)**:**
Um zu berechnen, wie wahrscheinlich es ist, dass es in zwei Tagen sonnig ist, wenn es heute regnerisch ist, erstellen wir zunächst einen **Startvektor** π_0, der den heutigen Zustand (also den Zustand π am Tag 0) beschreibt. In diesem Fall ist der Startvektor:

$$\pi_0 = \begin{pmatrix} 0 & 1 \\ \Uparrow & \Uparrow \\ \text{☀} & \text{☂} \end{pmatrix}$$

Nun multiplizieren wir den Startvektor π_0 mit der Übergangsmatrix P, um π_1, also den Zustand nach einem Tag, zu berechnen. Dabei verwenden wir die Matrixmultiplikation. Die Matrixmultiplikation ist definiert als:

$$\begin{pmatrix} a & b \end{pmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a \cdot e + b \cdot g & a \cdot f + b \cdot h \end{pmatrix}$$

Wir berechnen nun also π_1:

$$\pi_1 = \pi_0 \cdot P = \begin{pmatrix} 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.4 & 0.6 \end{pmatrix}$$

Die Einträge im Vektor π_1 geben die Wahrscheinlichkeiten für die Zustände nach einem Tag an. Die Wahrscheinlichkeit, dass es morgen sonnig ist, beträgt also 0.4. Um die Wahrscheinlichkeit zu berechnen, dass es in zwei Tagen sonnig ist, multiplizieren wir den Zustand nach einem Tag erneut mit der Übergangsmatrix:

$$\pi_2 = \pi_1 \cdot P = \begin{pmatrix} 0.4 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.56 & 0.44 \end{pmatrix}$$

Die Einträge in der Matrix π_2 geben die Wahrscheinlichkeiten für die Zustände nach zwei Tagen an. Die Wahrscheinlichkeit, dass es in zwei Tagen sonnig ist, beträgt also 0.56.

Aufgabe 3.4 App-Nutzung

Ein Schüler nutzt auf seinem Smartphone entweder Instagram (I) oder TikTok (T). Wenn er zuerst Instagram nutzt, so nutzt er mit einer Wahrscheinlichkeit von 0.6 wieder Instagram und mit 0.4 TikTok. Nach einer TikTok-Nutzung entscheidet er sich mit 0.1 für Instagram und mit 0.9 für TikTok. Die Übergangsmatrix P lautet:

		Nach I	T
Von	I	0.6	0.4
	T	0.1	0.9

Wie gross ist die Wahrscheinlichkeit, dass der Schüler **nach zwei Nutzungen bei Instagram landet, wenn er mit TikTok startet**? Berechnen Sie die Lösung, indem Sie den

Startvektor pi_0 erstellen und diesen mehrfach mit der Übergangsmatrix P multiplizieren.

✔ Lösungsvorschlag zu Aufgabe 3.4

Wir definieren den Startvektor π_0:

$$\pi_0 = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

Zuerst berechnen wir den Zustand nach einer Nutzung:

$$\pi_1 = \pi_0 \cdot P = \begin{pmatrix} 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.9 \end{pmatrix}$$

Dann berechnen wir den Zustand nach zwei Nutzungen:

$$\pi_2 = \pi_1 \cdot P = \begin{pmatrix} 0.1 & 0.9 \end{pmatrix} \cdot \begin{pmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{pmatrix} = \begin{pmatrix} 0.15 & 0.85 \end{pmatrix}$$

Die Wahrscheinlichkeit, dass der Schüler nach zwei Nutzungen bei Instagram landet, beträgt also 0.15.

Aufgabe 3.5 Pendlerverhalten

Eine Pendlerin fährt jeden Morgen entweder mit dem Fahrrad oder mit dem Bus zur Arbeit. Die Übergangswahrscheinlichkeiten sind:

- Nach einer Fahrt mit dem Fahrrad nimmt sie am nächsten Tag mit Wahrscheinlichkeit 0.9 wieder das Fahrrad und mit 0.1 den Bus.
- Nach einer Busfahrt nimmt sie am nächsten Tag mit Wahrscheinlichkeit 0.6 wieder den Bus und mit 0.4 das Fahrrad.

Wie sieht die Übergangsmatrix aus? Wie gross ist die Wahrscheinlichkeit, dass die Pendlerin nach zwei Tagen mit dem Bus fährt, wenn sie heute mit dem Fahrrad fährt?

✔ Lösungsvorschlag zu Aufgabe 3.5

Die Übergangsmatrix lautet:

$$P = \begin{array}{c|cc} & \text{Fahrrad} & \text{Bus} \\ \hline \text{Fahrrad} & 0.9 & 0.1 \\ \text{Bus} & 0.4 & 0.6 \end{array}$$

Der Startvektor ist:

$$\pi_0 = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

Wir berechnen den Zustand nach einem Tag:

$$\pi_1 = \pi_0 \cdot P = \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix}$$

Dann berechnen wir den Zustand nach zwei Tagen:

$$\pi_2 = \pi_1 \cdot P = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.85 & 0.15 \end{pmatrix}$$

Die Wahrscheinlichkeit, dass die Pendlerin nach zwei Tagen mit dem Bus fährt, wenn sie heute mit dem Fahrrad fährt, beträgt also 0.15.

Aufgabe (Challenge) 3.6 Tourist in Thailand

Die Übergangsmatrix für Aufgabe 3.3 lautet:

$$P = \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.4 & 0.3 & 0.3 \end{pmatrix}$$

Die Reihenfolge der Zustände ist: **F**estland, **S**amui, **P**hangan.

Wie gross ist die Wahrscheinlichkeit (in Prozent), dass der Tourist nach 3 Tagen wieder auf dem Festland ist, wenn er am ersten Tag auf dem Festland startet?

Die Matrixmultiplikation mit 3 Zuständen lautet:

$$\begin{pmatrix} a & b & c \end{pmatrix} \cdot \begin{pmatrix} e & f & g \\ h & i & j \\ k & l & m \end{pmatrix} = \begin{pmatrix} a \cdot e + b \cdot h + c \cdot k & a \cdot f + b \cdot i + c \cdot l & a \cdot g + b \cdot j + c \cdot m \end{pmatrix}$$

✔ Lösungsvorschlag zu Aufgabe 3.6

Wir berechnen die Verteilung nach 3 Tagen. Startvektor: $\pi_0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$.

Berechne $\pi_1 = \pi_0 \cdot P$:

$$\pi_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.4 & 0.3 & 0.3 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.7 & 0.2 \end{pmatrix}$$

Berechne $\pi_2 = \pi_1 \cdot P$:

$$\pi_2 = \begin{pmatrix} 0.1 & 0.7 & 0.2 \end{pmatrix} \cdot \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.4 & 0.3 & 0.3 \end{pmatrix}$$

Berechne jede Komponente:

$$\begin{aligned} \text{Festland:} \quad & 0.1 \cdot 0.1 + 0.7 \cdot 0.1 + 0.2 \cdot 0.4 = 0.01 + 0.07 + 0.08 = 0.16 \\ \text{Samui:} \quad & 0.1 \cdot 0.7 + 0.7 \cdot 0.4 + 0.2 \cdot 0.3 = 0.07 + 0.28 + 0.06 = 0.41 \\ \text{Phangan:} \quad & 0.1 \cdot 0.2 + 0.7 \cdot 0.5 + 0.2 \cdot 0.3 = 0.02 + 0.35 + 0.06 = 0.43 \end{aligned}$$

Also:

$$\pi_2 = \begin{pmatrix} 0.16 & 0.41 & 0.43 \end{pmatrix}$$

Berechne $\pi_3 = \pi_2 \cdot P$:

$$\pi_3 = \begin{pmatrix} 0.16 & 0.41 & 0.43 \end{pmatrix} \cdot \begin{pmatrix} 0.1 & 0.7 & 0.2 \\ 0.1 & 0.4 & 0.5 \\ 0.4 & 0.3 & 0.3 \end{pmatrix}$$

Berechne jede Komponente:

$$\begin{aligned} \text{Festland:} \quad & 0.16 \cdot 0.1 + 0.41 \cdot 0.1 + 0.43 \cdot 0.4 = 0.016 + 0.041 + 0.172 = 0.229 \\ \text{Samui:} \quad & 0.16 \cdot 0.7 + 0.41 \cdot 0.4 + 0.43 \cdot 0.3 = 0.112 + 0.164 + 0.129 = 0.405 \\ \text{Phangan:} \quad & 0.16 \cdot 0.2 + 0.41 \cdot 0.5 + 0.43 \cdot 0.3 = 0.032 + 0.205 + 0.129 = 0.366 \end{aligned}$$

Also:

$$\pi_3 = \begin{pmatrix} 0.229 & 0.405 & 0.366 \end{pmatrix}$$

Die Wahrscheinlichkeit, nach 3 Tagen auf dem Festland zu sein, beträgt 0.229, also etwa 23%.

3.5 Zufallssimulation einer Markov-Kette

Markov-Ketten lassen sich nicht nur rechnerisch mit Matrizen analysieren, sondern auch durch Zufallssimulation experimentell untersuchen. Dabei wird der Ablauf der Kette Schritt für Schritt mit Hilfe von Zufallszahlen simuliert. So kann man zum Beispiel typische Pfade, Häufigkeiten der Zustände oder das Langzeitverhalten beobachten.

3.5.1 Ablauf einer Simulation

Um eine Markov-Kette zu simulieren, gehen Sie wie folgt vor:

1. Wählen Sie einen beliebigen Startzustand und erstellen Sie den dazugehörigen Startvektor.
2. Erzeugen Sie für jeden Zeitschritt eine Zufallszahl zwischen 0 und 1.
3. Wechseln Sie gemäss der Übergangswahrscheinlichkeiten in den nächsten Zustand (s. Beispiel untenan).
4. Wiederholen Sie die Schritte, um einen Pfad der Kette zu erzeugen.

Beispiel 3.4 (Simulation eines Wettermodells)**:**
Angenommen, wir simulieren das Wettermodell aus dem Kapitelanfang mit den Zuständen *sonnig* und *regnerisch* und folgender Übergangsmatrix:

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$$

Wir starten z. B. mit *regnerisch*. Für jeden Tag ziehen wir eine Zufallszahl z:

- Ist es heute regnerisch: $z < 0.4 \rightarrow$ morgen sonnig, sonst regnerisch.
- Ist es heute sonnig: $z < 0.8 \rightarrow$ morgen sonnig, sonst regnerisch.

3.5.2 Python-Code zur Simulation

Mit Python lässt sich eine solche Simulation einfach umsetzen:

```
import random

# Anfangszustand
zustand = "Sonne"

# Liste der simulierten Zustände (wird in der Schleife fortlaufend erweitert)
history = [zustand]

# Anzahl Simulationsschritte
schritte = 20

for i in range(schritte):
    # Zufällig nächsten Zustand bestimmen (mit Übergangswahrscheinlichkeiten)
    if zustand == "Sonne":
        # Übergangswahrscheinlichkeiten vom Zustand "Sonne"
        wahrscheinlichkeit = P[0]
    else:
        # Übergangswahrscheinlichkeiten vom Zustand "Regen"
        wahrscheinlichkeit = P[1]
```

```
    # generiere nächsten simulierten Zustand mit Übergangswahrscheinlichkeiten
    zustand = random.choices(["Sonne", "Regen"], weights=wahrscheinlichkeit).pop()

    # Füge simulierten Zustand zur Liste hinzu
    history.append(zustand)

print(history)
```

Programm 3.1: Simulation einer Markov-Kette (Wettermodell)

Aufgabe 3.7

Simulieren Sie das Wettermodell 1000 Schritte lang und bestimmen Sie den Anteil der sonnigen Tage.

3.6 Lernziele

- ☐ Ich kann **Anwendungen** von Markov-Ketten in der Informatik beschreiben.
- ☐ Ich kann erklären, was eine **Markov-Kette** ist und wie die **Markov-Eigenschaft** funktioniert.
- ☐ Ich kann aus einer realen Problemstellung eine Abbildung in Form eines **gerichteten Graphen** erstellen.
- ☐ Ich kann aus einer realen Problemstellung eine **Übergangsmatrix** erstellen.
- ☐ Ich kann einen Zustand einer Markov-Kette mithilfe einer **Matrixmultiplikation** vorhersagen.
- ☐ Ich kann Markov-Ketten **in Python** implementieren
- ☐ Ich kann mithilfe einer Simulation berechnen, wie sich eine Markov-Kette **über mehrere Schritte entwickelt**.
- ☐ Ich kann die langfristige Verteilung von Markov-Ketten mit Simulationen **in Python** berechnen.

Anhang A

Herleitung der Methode der kleinsten Quadrate

Definition der Zielfunktion

Das Ziel der linearen Regression besteht darin, die bestmögliche Gerade $y = mx + q$ zu finden, die die Summe der quadrierten Residuen zwischen den vorhergesagten Werten und den tatsächlichen Datenpunkten minimiert. Die Summe der quadrierten Abstände zwischen den Datenpunkten und der Geraden wird definiert als:

$$S = \sum_{i=1}^{n} (y_i - (mx_i + q))^2$$

Partielle Ableitungen

Um S zu minimieren, nehmen wir die partiellen Ableitungen bezüglich m und q und setzen diese gleich Null.

Partielle Ableitung nach m

$$\frac{\partial S}{\partial m} = \frac{\partial}{\partial m} \sum_{i=1}^{n} (y_i - mx_i - q)^2$$

$$= \sum_{i=1}^{n} 2(y_i - mx_i - q)(-x_i)$$

$$= -2 \sum_{i=1}^{n} x_i (y_i - mx_i - q)$$

Auf Null setzen für die Minimierung:

$$\sum x_i y_i - m \sum x_i^2 - q \sum x_i = 0$$

$$\sum x_i y_i = m \sum x_i^2 + q \sum x_i \quad \text{(Gleichung 1)}$$

Partielle Ableitung nach q

$$\frac{\partial S}{\partial q} = \frac{\partial}{\partial q} \sum_{i=1}^{n} (y_i - mx_i - q)^2$$

$$= \sum_{i=1}^{n} 2(y_i - mx_i - q)(-1)$$

$$= -2 \sum_{i=1}^{n} (y_i - mx_i - q)$$

Auf Null setzen für die Minimierung:

$$\sum y_i - m \sum x_i - nq = 0$$

$$\sum y_i = m \sum x_i + nq \quad \text{(Gleichung 2)}$$

Lösung der Gleichungen

Aus Gleichung 2:

$$q = \frac{\sum y_i - m \sum x_i}{n} \quad \text{(Gleichung 3)}$$

Einsetzen der Gleichung 3 in Gleichung 1:

$$\sum x_i y_i = m \sum x_i^2 + \left(\frac{\sum y_i - m \sum x_i}{n} \right) \sum x_i$$

$$\sum x_i y_i = m \sum x_i^2 + \frac{(\sum y_i) \sum x_i}{n} - \frac{m(\sum x_i)^2}{n}$$

Auflösen nach m:

$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Bestimmen von q über Einsetzen von m in Gleichung 3:

$$q = \frac{\sum y_i - m \sum x_i}{n}$$

Anhang B

Formelle Definition von Markov-Ketten

Formal definieren wir eine Markov-Kette über einen diskreten Zustandsraum S und eine Übergangsmatrix P.

Definition B.1 (Markov-Kette)**:**
Eine **Markov-Kette** ist ein stochastischer Prozess $\{X_n : n \in \mathbb{N}_0\}$ mit Werten in einem abzählbaren Zustandsraum S, der die Markov-Eigenschaft erfüllt:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

Die Wahrscheinlichkeit $p_{ij} = P(X_{n+1} = j | X_n = i)$ bezeichnet die Übergangswahrscheinlichkeit von Zustand i zu Zustand j. Die Matrix $P = (p_{ij})_{i,j \in S}$ heisst **Übergangsmatrix** der Markov-Kette.
Eine Markov-Kette heisst **homogen**, wenn die Übergangswahrscheinlichkeiten nicht vom Zeitpunkt n abhängen, sondern nur von den beteiligten Zuständen i und j.

Der Zustandsvektor $\pi^{(n)} = (\pi_1^{(n)}, \pi_2^{(n)}, \ldots)$ gibt die Wahrscheinlichkeitsverteilung der Zustände zum Zeitpunkt n an, wobei $\pi_i^{(n)} = P(X_n = i)$. Der Anfangszustand wird durch den Vektor $\pi^{(0)}$ beschrieben.

Die zeitliche Entwicklung einer Markov-Kette lässt sich durch wiederholte Multiplikation mit der Übergangsmatrix P berechnen:

$$\pi^{(n+1)} = \pi^{(n)} \cdot P$$

Das bedeutet, dass die Verteilung nach n Schritten durch $\pi^{(n)} = \pi^{(0)} \cdot P^n$ gegeben ist.

Data Science Cheatsheets

Das folgende Cheatsheet enthält die wichtigsten Befehle für den Umgang mit den Libraries `pandas`, `matplotlib` und `statsmodels`.

Daten analysieren und visualisieren

Die Libraries `pandas` und `matplotlib` sind die wichtigsten Werkzeuge für die Datenanalyse und -visualisierung in Python. `statsmodels` wird für statistische Modelle verwendet.

Libraries Importieren

```
import pandas as pd # für Datenanalyse, Tabellen
import matplotlib.pyplot as plt # für Grafiken
import statsmodels.formula.api as smf # für
    statistische Modelle
import numpy as np # für numerische Berechnungen
```

CSV-Datei laden

```
df = pd.read_csv('datei.csv')
```

Erste/letzte Zeilen anzeigen

```
df.head()   # erste 5 Zeilen
df.tail()   # letzte 5 Zeilen
```

Typ einer Variable anzeigen

```
type(df) # Tabelle (DataFrame)
type(df['col']) # Spalte (Series)
type(1) # Zahl (int)
type('Text') # Text (str)
df.dtypes # Daten-Typen aller Spalten
```

Spalten auswählen

```
df['Spalte']   # einzelne Spalte
df[['A', 'B']] # mehrere Spalten
```

Tabellen-Statistiken

```
df['Spalte'].mean()  # Mittelwert
df['Spalte'].sum()   # Summe
df['Spalte'].min()   # Minimum
df['Spalte'].max()   # Maximum
df['Spalte'].count() # Anzahl Einträge
df['Spalte'].std()   # Standardabweichung
df['Spalte'].unique() # Einzigartige Werte
df['Spalte'].value_counts() # Häufigkeit der Werte
df['Spalte'].shape # Anzahl Zeilen und Spalten
```

Zeilen filtern

```
df[df['Alter'] > 18] # Filter
```

Neue Spalte erstellen

```
df['BMI'] = df['Gewicht'] / (df['Grösse']/100)**2
```

Grafiken erstellen

```
df["Spalte"].value_counts().plot.pie() # Kuchendiagramm
df["Spalte"].value_counts().plot.bar() # Balkendiagramm
df.plot.line(x='x_spalte', y='y_spalte') #
    Liniendiagramm
df.plot.scatter(x='x_spalte', y='y_spalte') #
    Streudiagramm

# scatter plots nach Kategorien einfärben
df['kategorie_spalte'] = df['kategorie_spalte'].astype(
    'category')
df.plot.scatter(x='x_spalte', y='y_spalte', c='
    category_spalte', colormap='viridis')

# Histogramm mit 10 Grössenklassen
df['Spalte'].plot.hist(bins=10)
```

Gruppieren

Nach `spalte_gruppe` gruppieren und Mittelwert von `spalte_statistik` berechnen:

```
df.groupby('spalte_gruppe')["spalte_statistik"].mean()
```

Lineare Regression

Die folgenden Boxen enthalten die wichtigsten Formeln und Befehle für die lineare Regression von Hand und in Python.

Lineare Funktion

Formel einer linearen Funktion:

$$y = m \cdot x + q$$

Wobei m die Steigung und q den y-Achsenabschnitt darstellt.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$
$$q = y_1 - m \cdot x_1$$

Lineare Regression

- **Unabhängige Variable**: `df['Schlaf']`
- **Abhängige Variable**: `df['Note']`

```
model = smf.ols(formula='Note ~ Schlaf', data=df).fit()
print(model.params) # Koeffizienten

# Vorhersage
df['Note_pred'] = model.predict(df['Schlaf'])
```

Markov-Ketten

Die folgenden Boxen enthalten die wichtigsten Formeln und Befehle für Markov-Ketten, von Hand und in Python.

Markov-Ketten: Matrix-Multiplikation

$$\begin{pmatrix} a & b \end{pmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a \cdot e + b \cdot g & a \cdot f + b \cdot h \end{pmatrix}$$

Markov-Ketten: Übergangsmatrix erstellen

```
# Zustände: 0 = Sonne, 1 = Regen
P = np.array([
    [0.9, 0.1],
    [0.2, 0.8]
])
```

Markov-Ketten: Vorhersage (Matrixmultiplikation)

```
# Startzustand: 100% Sonne
pi0 = np.array([1, 0])

# Nach einem Tag
pi1 = pi0 @ P
print("Nach 1 Tag:", pi1)

# Nach zwei Tagen
pi2 = pi1 @ P
print("Nach 2 Tagen:", pi2)
```

Markov-Ketten: Vorhersage (Zufallsimulation)

```
zustand = "Sonne"     # Anfangszustand
history = [zustand]   # Liste simulierter Zustände
schritte = 20         # Anzahl Simulationsschritte

for i in range(schritte):
    if zustand == "Sonne":
        wahrscheinlichkeit = P[0]
    else:
        wahrscheinlichkeit = P[1]

    zustand = random.choices(["Sonne", "Regen"],
     weights=wahrscheinlichkeit).pop()
    history.append(zustand)

print(history)
```