Bausteine der Datenanalyse

Methodenbausteine Statistik

Lukas Arnold Simone Arnold Florian Bagemihl
Matthias Baitsch Marc Fehr Franca Hollmann
Maik Poetzsch Sebastian Seipel

2026-03-19

Untersuchung von zwei Merkmalen

Eine wichtige Anwendung statistischer Methoden besteht darin zu untersuchen, wie zwei Merkmale zusammenhängen. Zum Beispiel kann man sich fragen:

- Jedes Jahr geben Expert:innen eine Prognose über das erwartete Wirtschaftswachstum im kommenden Jahr ab. Wie gut passen Prognose und Wirklichkeit zusammen?
- Die mittlere Lebenserwartung von Menschen variiert in den Ländern der Welt deutlich. Gibt es einen Zusammenhang mit wirtschaftlichen Kenngrößen?

Im folgenden Abschnitt werden wir Methoden kennenlernen, mit denen solche Zusammenhänge dargestellt und berechnet werden können. Wir konzentrieren uns dabei auf metrische Merkmale und den zweidimensionalen Fall. Das heißt, wir betrachten Datensätze der Form (x_i, y_i), $i = 1, \dots, n$ wobei $x_i, y_i \in \mathbb{R}$ vorliegende Zahlenwerte sind.

Prognose des Sachverständigenrats: Der am statistischen Bundesamt angesiedelte Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung (die fünf Wirtschaftsweisen) veröffentlicht seit 1963 jährlich eine Prognose zum Wirtschaftswachstum des nächsten Jahres. Die Prognosen und das tatsächliche Wachstum sind für die Jahr 1975 bis 1997 in der folgenden Tabelle zusammengestellt.

	1975	1976	1977	1978	1979	1980	1981	1982
Prognose	2.0	4.5	4.5	3.5	3.75	2.75	0.5	0.5
Wachstum	-3.6	5.6	2.4	3.4	4.40	1.80	-0.3	-1.2

	1983	1984	1985	1986	1987	1988	1989	1990
Prognose	1.0	2.5	3.0	3.0	2.0	1.5	2.5	3.0
Wachstum	1.2	2.6	2.5	2.5	1.7	3.4	4.0	4.6

	1991	1992	1993	1994	1995	1996	1997
Prognose	3.5	2.5	0.0	0.0	3.0	2.0	2.5
Wachstum	3.4	1.5	-1.9	2.3	1.9	1.4	2.2

Es liegt in der Natur von Prognosen, dass sie von der später beobachteten Wirklichkeit abweichen. Interessant ist es jedoch zu fragen, in welchem Grad die Vorhersagen zutreffen. Ein Stück weit lässt sich bereits an der Tabelle ablesen, dass die Prognosen in vielen Fällen die wirtschaftliche Entwicklung gut vorhergesagt haben. Es gibt aber auch Vorhersagen, die weit danebenlagen, zum Beispiel in den Jahren 1975 oder 1982.

Mittlere Lebenserwartung: Mit den Daten der Weltbank betrachten wir den Zusammenhang zwischen der mittleren Lebenserwartung in einem Land und

- dem Bruttoinlandsprodukt (BIP),
- den Ausgaben für das Gesundheitswesen (in % des BIP),
- den Ausgaben für Bildung (in % des BIP) sowie
- dem Gini-Koeffizienten.

Welcher dieser Indikatoren hängt am stärksten mit der Lebenserwartung zusammen?

Graphische Darstellung

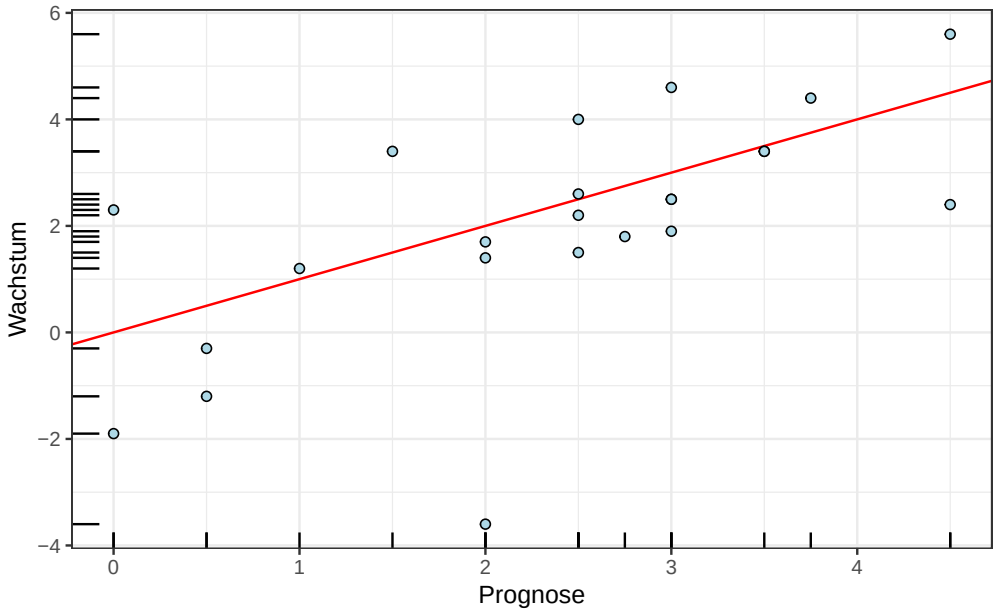
Streudiagramm

Die einfachste Möglichkeit zwei gemeinsame Messwerte $(x_i, y_i), i = 1, \dots, n$ darzustellen ist das so genannte Streudiagramm, das gegebenenfalls zu einem Blasendiagramm erweitert werden kann.

Definition 0.1 (Streudiagramm)**.** Die Darstellung der Wertepaare $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ durch geometrische Objekte (Punkte, Kreise) in einem xy-Koordinatensystem heißt Streudiagramm. Werden die Symbole nach einem dritten Merkmal skaliert, dann spricht man von einem Blasendiagramm.

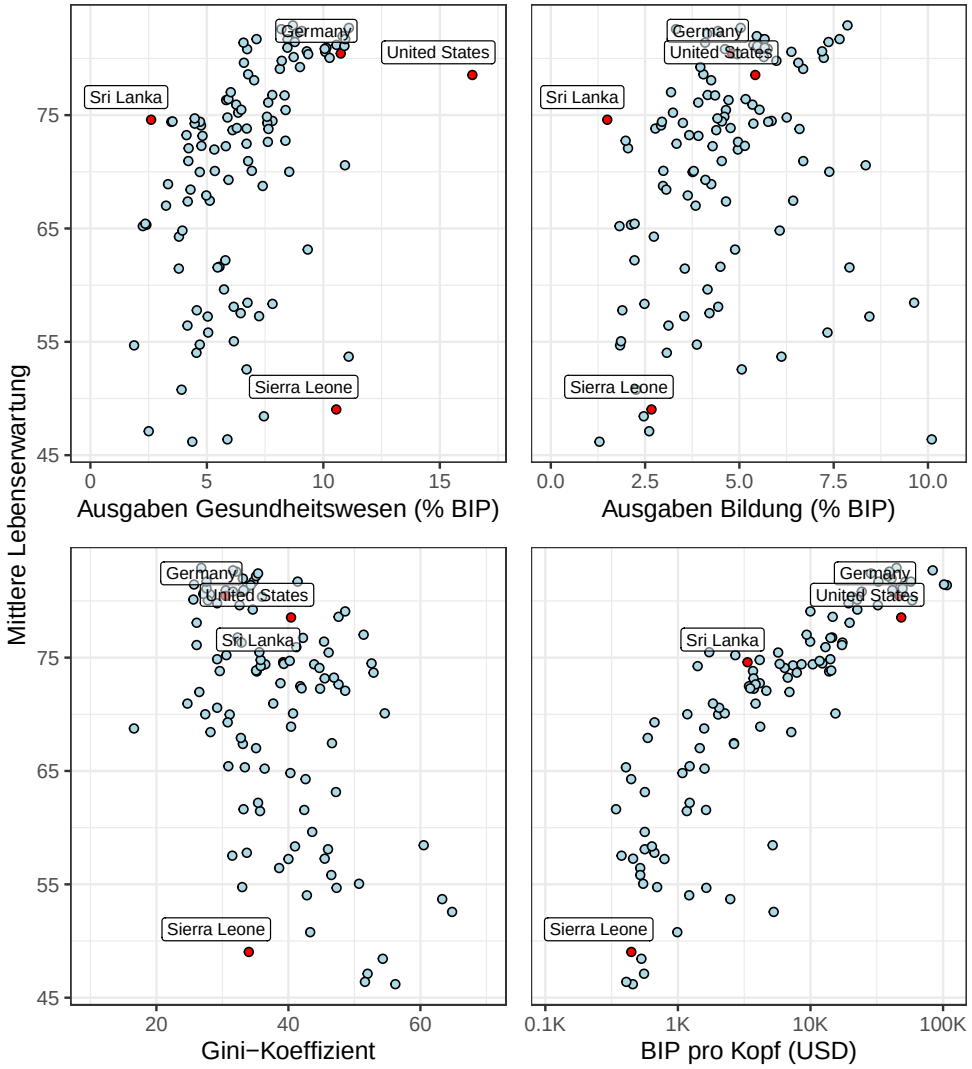
Sachverständigenrat

Unten dargestellt ist das Streudiagramm des tatsächlichen Wirtschaftswachstums geplottet über die Vorhersagen des Sachverständigenrats. Dabei sind auf den jeweiligen Achsen noch durch kurze Striche die Lage der einzelnen Werte angedeutet. Könnte der Sachverständigenrat in die Zukunft sehen, lägen alle Punkte auf der eingezeichneten Winkelhalbierenden. In der Wirklichkeit ist das nicht der Fall und je schlechter die Prognose war, umso weiter sind die Punkte von dieser Diagonale entfernt.

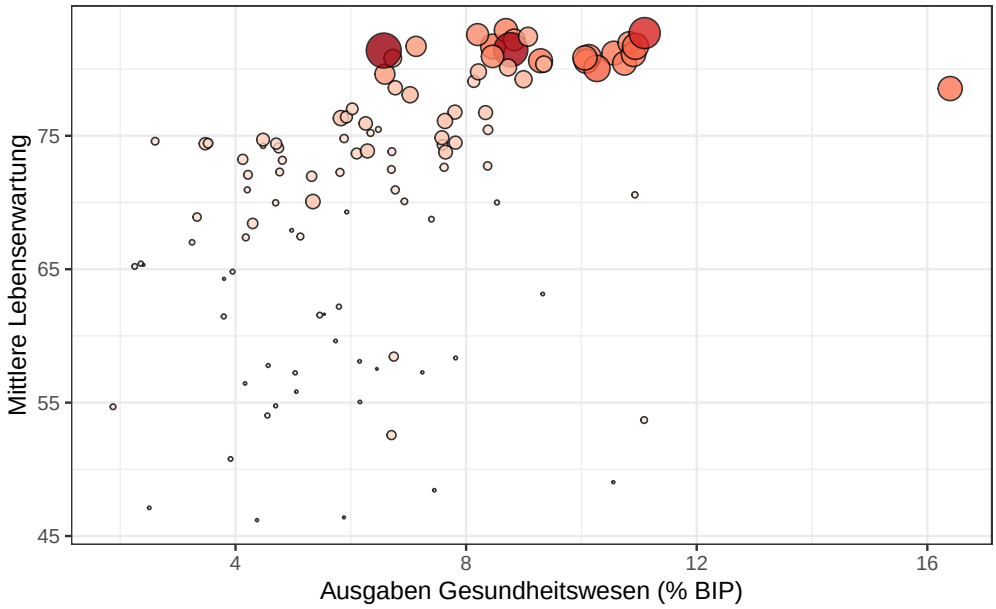


Lebenserwartung

In den folgenden Plots ist die mittlere Lebenserwartung über verschiedene andere Indikatoren geplottet. Beachten Sie, dass im Plot über das Bruttoinlandsprodukt eine logarithmische Skala verwendet wird.



Eine erweiterte Sicht lässt sich gewinnen, wenn in dem Plot der Lebenserwartung über die prozentualen Ausgaben im Gesundheitswesen noch das BIP pro Kopf berücksichtigt wird. Dies führt zu dem dargestellten Blasendiagramm, in dem die Wirtschaftsleitung der Länder auf Größe und Farbe der Kreise abgebildet ist.



Zweidimensionale Histogramme und Dichten

Insbesondere wenn es sich um umfangreiche Datensätze handelt, überdecken sich in einem Streuplot häufig die Punkte. Diese Problematik lässt sich mithilfe von zweidimensionalen Histogrammen oder zweidimensionalen Kerndichteschätzern umgehen.

Für das zweidimensionale Histogramm definiert man Klassen für beide zu betrachtende Merkmale X und Y

$$[c_0^x, c_1^x), [c_1^x, c_2^x), \ldots, [c_{k-1}^x, c_k^x) \quad \text{und} \quad [c_0^y, c_1^y), [c_1^y, c_2^y), \ldots, [c_{m-1}^y, c_m^y)$$

und zählt für jedes Rechteck $[c_{i-1}^x, c_i^x) \times [c_{j-1}^y, c_j^y)$ die Punkte, die in diesem Rechteck liegen. Damit ergeben sich die Klassenhäufigkeiten h_{ij}. Über jedem Rechteck ij wird dann ein Quader angeordnet, dessen Volumen gleich der zugehörigen Häufigkeit h_{ij} beziehungsweise der relativen Häufigkeit h_{ij}/n ist.

Alternativ kann auch im zweidimensionalen Fall ein Kerndichteschätzer verwendet werden. Dieser funktioniert im Prinzip genauso wie für eine Variable, nur dass wir jetzt das Produkt von zwei Kernfunktionen verwenden:

$$\hat{f}(x, y) = \frac{1}{n\, h_1 h_2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_1}\right) K\left(\frac{y - y_i}{h_2}\right).$$

Die Glättungseigenschaften des Schätzers hängen von den Parametern h_1 und h_2 sowie der Wahl der Kernfunktion K ab (**?@sec-approximation-dichtekurven**).

In einer zweidimensionalen Darstellung des Histogramms oder der Dichtefunktion werden die Häufigkeiten farblich kodiert, siehe Abbildung 1. Eine weitere Möglichkeit besteht in einer dreidimensionalen Darstellung wie in Abbildung 2 zu sehen. Die ebene Darstellung ist mit ggplot erstellt, die räumliche Darstellung mit dem Programm Mathematica.

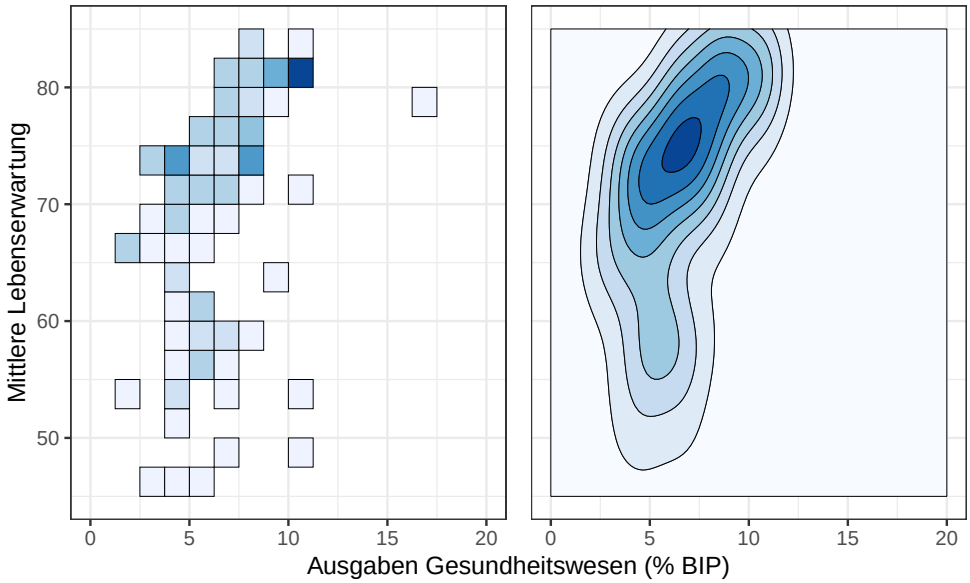


Abbildung 1: Histogramms (links) und geschätzte Verteilungsfunktion (rechts), ebene Darstellung

Zusammenhangsmaß: Empirischer Korrelationskoeffizient von Bravais und Pearson

Mit der graphischen Darstellung in einem Streudiagramm oder einer Dichteschätzung lässt sich die Anordnung von Beobachtungen zu zwei Merkmalen veranschaulichen. Ist dabei zu sehen, dass mit wachsenden Werten des einen Merkmals die Werte des anderen Merkmals ebenfalls zunehmen (oder abnehmen), dann lässt sich vermuten, dass zwischen den beiden Merkmalen ein Zusammenhang besteht. Mit dem empirischen Korrelationskoeffizienten kann gemessen werden, inwieweit es sich dabei um einen linearen Zusammenhang handelt.

Es geht nun darum, für die Beobachtungen $x_1, x_2, \ldots x_n$ und $y_1, y_2, \ldots, y_n$ zu zwei metrischen Merkmalen X und Y eine Zahl r zu bestimmen, mit der die Stärke des linearen Zusammenhangs zwischen zwei Merkmalen gemessen wird. Dabei soll r einen positiven Wert annehmen, wenn die Punkte nahe einer Geraden mit positiver Steigung liegen. Fällt die Gerade hingegen, dann soll der Wert negativ werden. Besteht kein Zusammenhang zwischen den Merkmalen oder lässt sich der Zusammenhang nicht durch eine lineare Funktion darstellen, dann soll r nahe Null liegen. Diese Eigenschaften des gesuchten Koeffizienten r sind in Abbildung 3 dargestellt.

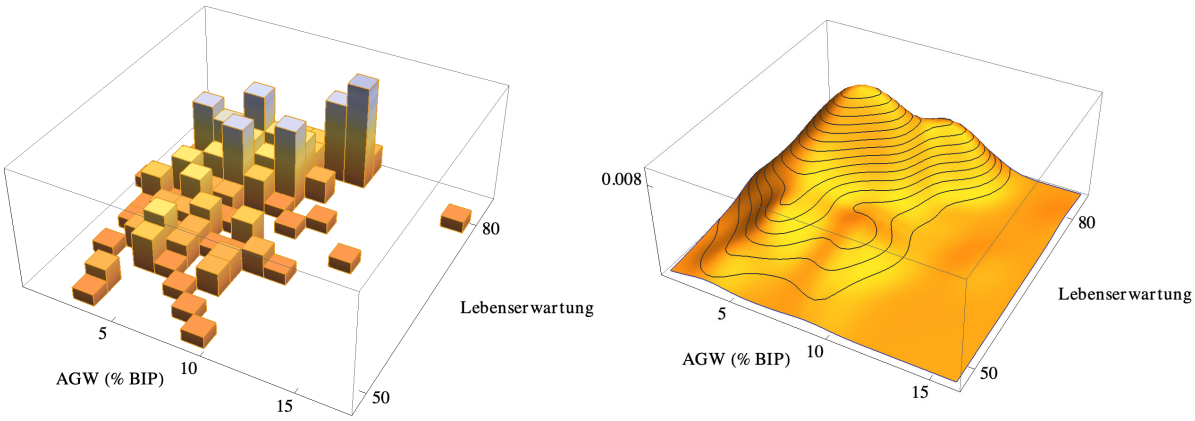


Abbildung 2: Histogramms (links) und geschätzte Verteilungsfunktion (rechts), räumliche Darstellung

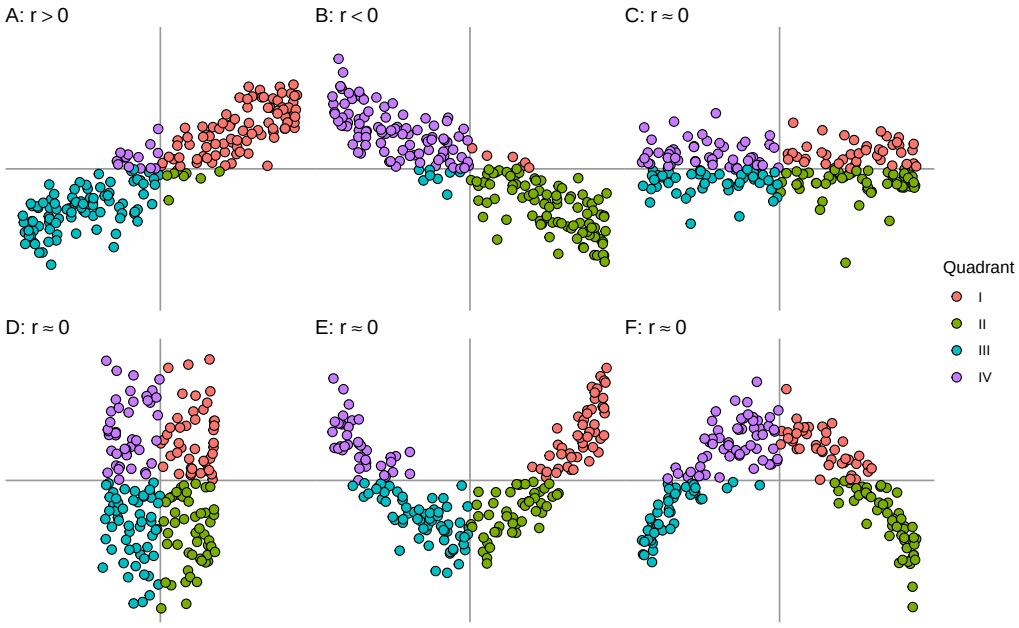


Abbildung 3: Mögliche Situationen für den Korrelationskoeffizienten r

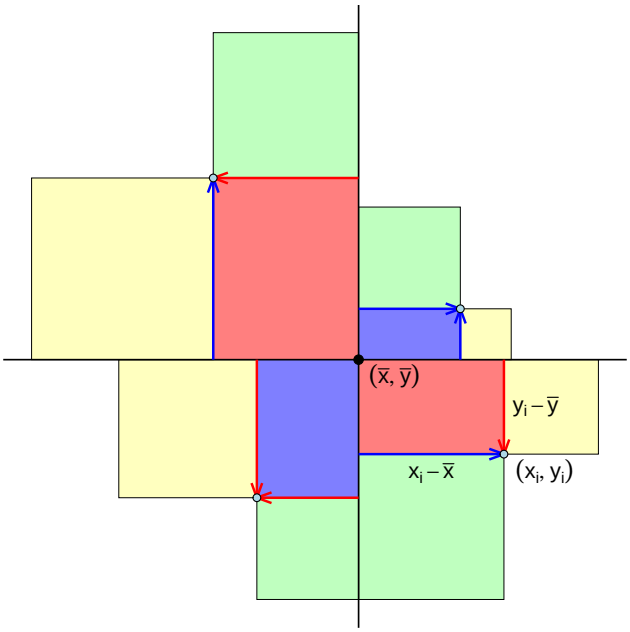
Wie lässt sich nun eine solche Zahl bestimmen? Die grundlegende Idee besteht darin, dass im Streudiagramm ein Koordinatensystem in das Zentrum der Punktwolke gelegt wird. Für jeden Punkt (x_i, y_i) lässt sich dann überprüfen, in welchem Quadrant des neuen Koordinatensystems der Punkt liegt. In Abbildung 3 ist dieser Zusammenhang farblich hervorgehoben.

Fall A: Die überwiegende Zahl der Punkte liegt in den Quadranten I und III

Fall B: Hier liegen die allermeisten Punkte in den Quadranten II und IV

Fälle C – F: Die Punkte verteilen sich auf alle Quadranten

Für die Lage des neuen Koordinatensystems verwenden wir die Mittelwerte $\bar{x}$ und $\bar{y}$ der Merkmale X und Y. Damit können wir uns nun überlegen, wie der Korrelationskoeffizient bestimmt werden kann.



Um die Lage der Punkte (x_i,y_i) im jeweiligen Quadranten zu berücksichtigen, werden für $i = 1, ..., n$ die (orientierten) Flächeninhalte der Rechtecke mit den Seitenlängen $x_i - \bar{x}$ und $y_i - \bar{y}$ zu Grunde gelegt. In der Summe ergibt sich damit, wie es sein soll, in den Fällen A und B eine positive bzw. negative Zahl. Für die Fälle C bis D werden sich die Flächeninhalte in etwa aufheben. Zusätzlich wird, um den Wertebereich des Korrelationskoeffizienten auf das Intervall $[-1, 1]$ einzuschränken, noch durch die Wurzel des Produktes der Summen der Abstandsquadrate $(x_i - \bar{x})^2$ (grün) und $(y_i - \bar{y})^2$ (gelb) dividiert. Wir erhalten damit den Korrelationskoeffizienten

$$r = \frac{\text{Summe der blauen und roten Flächen}}{\sqrt{(\text{Summe der gelben Flächen}) \cdot (\text{Summe der grünen Flächen})}},$$

den wir nochmal formal festhalten wollen.

Definition 0.2 (Bravais-Pearson-Korrelationskoeffizient)**.** Für die Werte $(x_i, y_i), i = 1, \dots, n$ heißt die Zahl

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Bravais-Person-Korrelationskoeffizient . Der Wertebereich ist $-1 \leq r \leq 1$.

Je nachdem ob der Korrelationskoeffizient positiv oder negativ ist, liegt eine der folgende Situationen vor.

Falls $r > 0$: Positive Korrelation, gleichsinniger linearer Zusammenhang. Die Werte (x_i, y_i) liegen tendenziell um eine Gerade positiver Steigung.

Falls $r < 0$: Negative Korrelation, gegensinniger linearer Zusammenhang. Die Werte (x_i, y_i) liegen tendenziell um eine Gerade negativer Steigung.

Entsprechend dem Zahlenwert von r unterscheidet man grob die Korrelationsgrade

Für $r \approx 0$: Keine Korrelation, unkorreliert, kein linearer Zusammenhang.

Für $|r| < 0.5$: Schwache Korrelation.

Für $0.5 \leq |r| < 0.8$: Mittlere Korrelation.

Für $0.8 \leq |r| < 1$: Starke Korrelation.

Für $|r| = 1$: Die Werte liegen exakt auf einer Geraden.

Neben dem Korrelationskoeffizienten von Bravais und Pearson gibt es noch weitere Zusammenhangsmaße, zum Beispiel den Spaerman-Korrelationskoeffizienten. Diese sind aber weniger gebräuchlich und werden hier nicht weiter diskutiert.

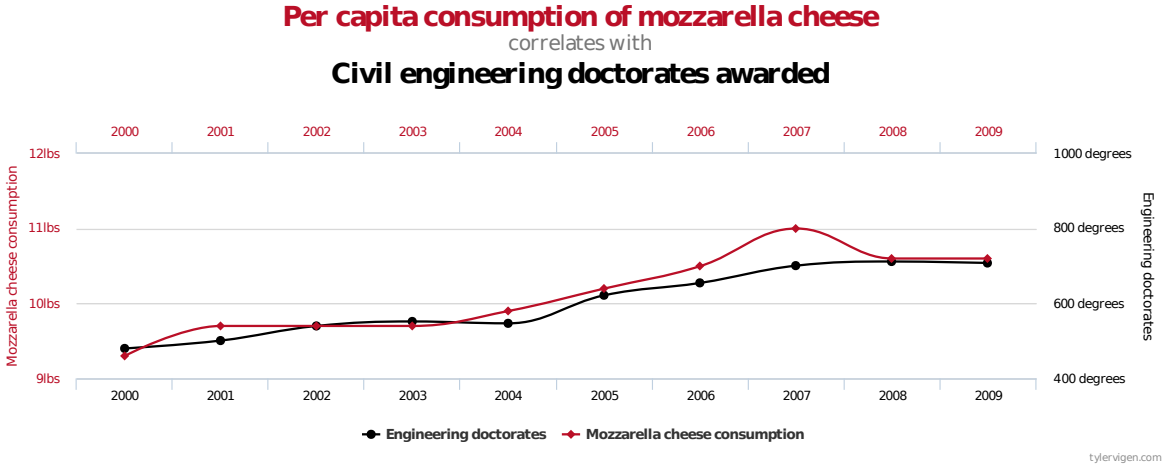
Korrelation und Kausalität

In der Gleichung für den Bravais-Pearson-Korrelationskoeffizienten werden die Werte beider Merkmale gleich behandelt. Der Wert von r gibt also keinerlei Auskunft darüber, in welcher Richtung eine mögliche Beeinflussung zwischen den Merkmalen stattfindet. Der Zahlenwert von r darf daher nie losgelöst von dem sachlogischen Zusammenhang betrachtet werden.

Ein weiteres Problem besteht darin, dass es häufig verlockend ist, einen hohen Korrelationswert vorschnell als kausalen Zusammenhang zu interpretieren. Kausalzusammenhänge können niemals allein durch große Werte eines Zusammenhangmaßes begründet werden. Um einen kausalen

Zusammenhang zu begründen, vielmehr muss stets ein Zusammenhang zwischen Ursache und Wirkung gefunden werden, der sich ohne Bezugnahme auf die statistischen Werte begründen lässt.

Die Webseite http://tylervigen.com (mittlerweile auch als Buch erhältlich) verdeutlicht das an einer Reihe kurioser Beispiele. Dort sind eine Vielzahl von Daten aufgelistet, die einen sehr hohen Korrelationsgrad aufweisen, aber offensichtlich völlig unabhängig voneinander sind. Zum Beispiel weist in den USA der Konsum von Mozarella und die Anzahl von Promotionen in Ingenieursfächern in den Jahren 2000 bis 2009 einen Korrelationskoeffizienten von $r = 0.95$ auf.

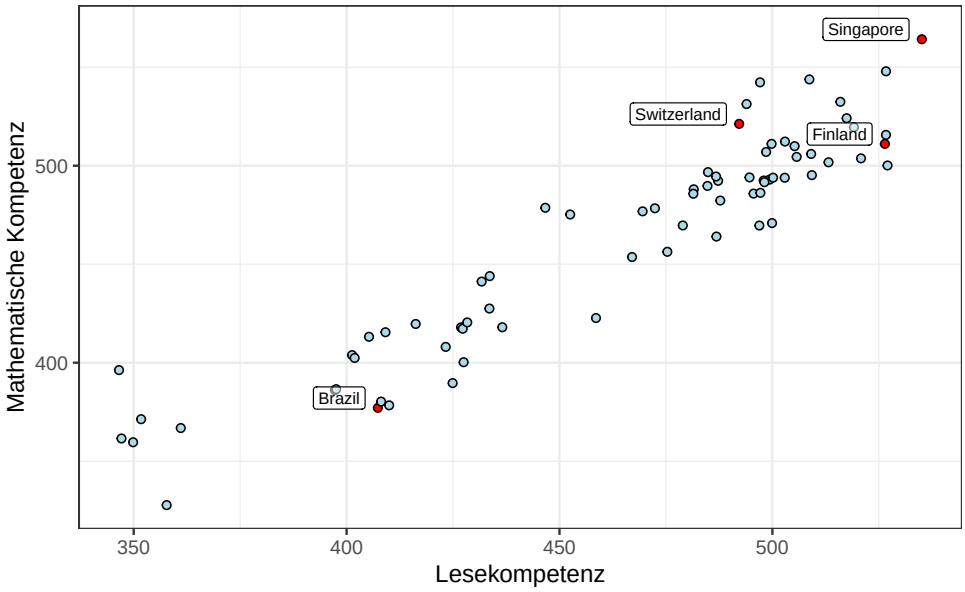


Wir merken uns daher:

$$\text{Korrelation} \quad \neq \quad \text{Kausalität}$$

Man sollte eine hohe Korrelation lediglich als Hinweis auf einen möglichen kausalen Zusammenhang verstehen, der näher zu untersuchen ist. Dazu sind stets sachlogische Überlegungen erforderlich. Zusätzlich ist immer zu bedenken, dass weitere wesentliche Merkmale unter Umständen übersehen wurden. Dies kann zu so genannten Scheinkorrelationen oder auch zu verdeckten Korrelationen führen. (Fahrmeir u. a. 2016)

> **Beispiel Scheinkorrelation:** In der PISA-Studie der OECD wurden zuletzt im Jahr 2015 die Ergebnisse einer weltweiten Studie zu den Fähigkeiten von Schülern in verschiedenen Kompetenzbereichen veröffentlicht. Die Rohdaten der Studie stehen unter http://www.oecd.org/pisa/data zum Download bereit. Der Datensatz umfasst 519334 Beobachtungen von 921 Merkmalen, das entsprechende Datenfile ist ca. 1.5 GB groß.



Dargestellt sind die Ergebnisse in den Kompetenzbereichen Mathematik und Lesen. Offenbar sind die Ergebnisse für beide Kompetenzbereiche eng miteinander korreliert. Dies spiegelt sich in dem Zahlenwert von $r = 0.94$ wider. Allerdings ist die Schlussfolgerung, dass eine gute Lesekompetenz automatisch auch eine gute Kompetenz im mathematischen Bereich zur Folge hat (oder umgekehrt), sicherlich nicht die passende Erklärung. Vielmehr muss man hier nach Einflussfaktoren fragen, die im Schulsystem und im kulturellen Umfeld angesiedelt sind.

Lineare Regression

Mithilfe der Korrelationsanalyse lassen sich ungerichtete Zusammenhänge zwischen zwei Merkmalen X und Y untersuchen. In vielen Fällen legt allerdings der untersuchte Sachverhalt nahe, dass eine der beobachteten Größen von der anderen Größe abhängt. Allerdings handelt es sich in der Regel nicht um einen rein deterministischen Zusammenhang der Form

$$Y = f(X),$$

sondern es kommt noch eine zufällige Abweichung hinzu. Wenn wir von einem additiven Fehlerterm ausgehen, dann erhalten wir die Beziehung

$$Y = f(X) + \varepsilon,$$

wobei ε eine zufällige Größe ist. Ziel bei der Bestimmung der Funktion f ist es dabei, einen möglichst großen Anteil an der Variabilität der Daten durch diese Funktion zu erklären. Eine

Beziehung dieser Art wird Regression genannt, der Funktion f fällt dabei die Rolle eines Regressionsmodells zu. In dieser Schreibweise verwenden wir die Großbuchstaben X und Y um zum Ausdruck zu bringen, dass wir über eine Eigenschaft des gesamten Datensatzes sprechen und nicht über einzelne Werte.

Die einfachste Möglichkeit ist es, für die Funktion f eine lineare Beziehung anzunehmen. Damit erhalten wir die Zuordnungsvorschrift

$$f(x) = \alpha + \beta x,$$

wobei α der y-Achsenabschnitt und β die Steigung der Geraden sind. Der Kleinbuchstabe x steht jetzt für eine beliebige Zahl. Wenn wir die Datenpaare $(x_i, y_i), i = 1, \dots, n$ in die Funktionsgleichung einsetzen, dann ergibt sich die empirische Beziehung

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

wobei ε_i den Fehler erfasst, der sich aus der Geradenanpassung ergibt. Der Zusammenhang ist in Abbildung 4 dargestellt.

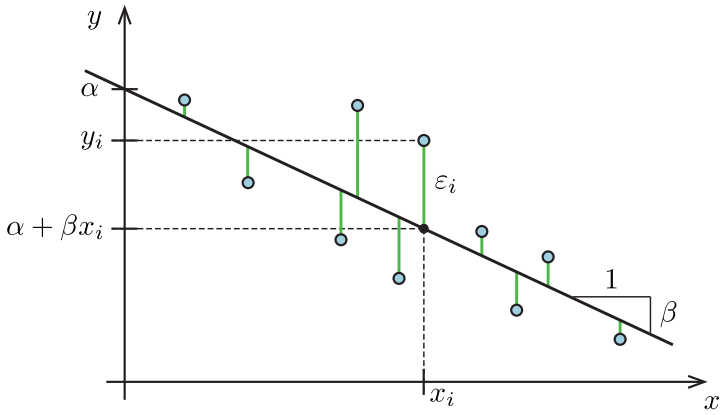


Abbildung 4: Ausgleichsgerade

Berechnung der Ausgleichsgeraden

Es stellt sich nun die Frage, wie die Parameter α und β der Ausgleichsgeraden gewählt werden sollen. Ziel ist es, die Abstände $\varepsilon_i = y_i - (\alpha + \beta x_i)$, also die Länge der grünen Linien in Abbildung 4 insgesamt möglichst klein werden zu lassen.

Natürlich könnten wir wieder wie beim Median, die Beträge der Abstände summieren. Allerdings ist die Betragsfunktion in der Anwendung unhandlich und wir summieren stattdessen die Abstandsquadrate. Gesucht sind also Zahlen α und β, so dass die Summe

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

möglichst klein wird. Die Werte von α und β, für die die Summe ihr Minimum annimmt, nennt man Kleinste-Quadrate-Schätzer und diese Methode zur Bestimmung der Ausgleichsgeraden die Methode der kleinsten Quadrate.

Zur Bestimmung der Kleinste-Quadrate-Schätzer stellen wir die Funktion $Q : R^2 \to R$ mit

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

auf. Im Minimum der Funktion müssen die partiellen Ableitungen

$$\frac{\partial Q(\alpha, \beta)}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))$$

und

$$\frac{\partial Q(\alpha, \beta)}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - (\alpha + \beta x_i))$$

beide gleich null sein. Wir formen die erste Gleichung um:

$$\begin{aligned} -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) = 0 \qquad & \iff \\ \sum_{i=1}^{n} y_i - \alpha n - \beta \sum_{i=1}^{n} x_i = 0 \qquad & \iff \\ \alpha = \bar{y} - \beta \bar{x}. & \end{aligned}$$

Aus der zweiten Gleichung erhalten wir

$$\begin{aligned} -2 \sum_{i=1}^{n} x_i (y_i - (\alpha + \beta x_i)) = 0 \quad & \iff \\ \sum_{i=1}^{n} x_i y_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0. & \end{aligned}$$

Einsetzen von $\alpha = \bar{y} - \beta \bar{x}$ und $\sum_{i=1}^{n} x_i = n \bar{x}$ liefert

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta \bar{x}) n \bar{x} - \beta \sum_{i=1}^{n} x_i^2 = 0 \quad & \iff \\ \sum_{i=1}^{n} x_i y_i - n\, \bar{x}\, \bar{y} - \beta \left(\sum_{i=1}^{n} x_i^2 - n\, \bar{x}^2 \right) = 0, & \end{aligned}$$

so dass

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\, \bar{x}\, \bar{y}}{\sum_{i=1}^{n} x_i^2 - n\, \bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

gelten muss. Für die letzte Umformung wurden die Beziehungen

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

und

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

verwendet, die sich in einer einfachen Nebenrechnung nachweisen lassen. Zusammengefasst ergibt sich folgende Definition.

Definition 0.3 (Lineare Regression und Kleinste-Quadrate-Schätzer)**.** Für die Werte $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ heißt der Zusammenhang

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

lineare Einfachregression. Dabei ist α der y-Achsenabschnitt, β die Steigung und ε_i der Fehler.

Die Kleinste-Quadrate-Schätzer

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{und} \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

minimiert die Summe der Fehlerquadrate.

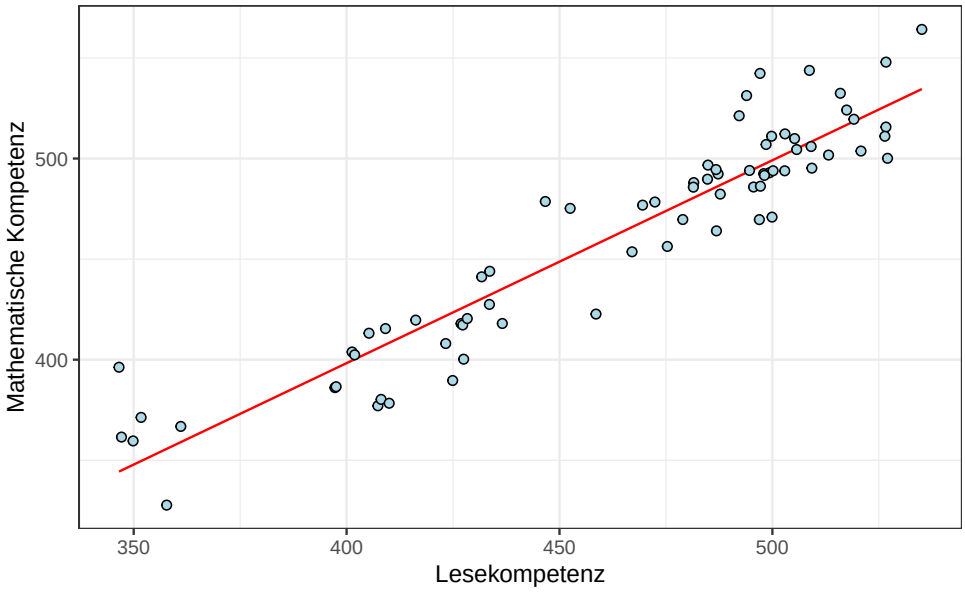
Beispiel PISA-Studie: Für die Daten zur Lesekompetenz X und der mathematischen Kompetenz Y aus der PISA-Studie erhalten wir mit den Mittelwerten

$$\bar{x} = 461.8 \quad \text{und} \quad \bar{y} = 463.0$$

die Parameter

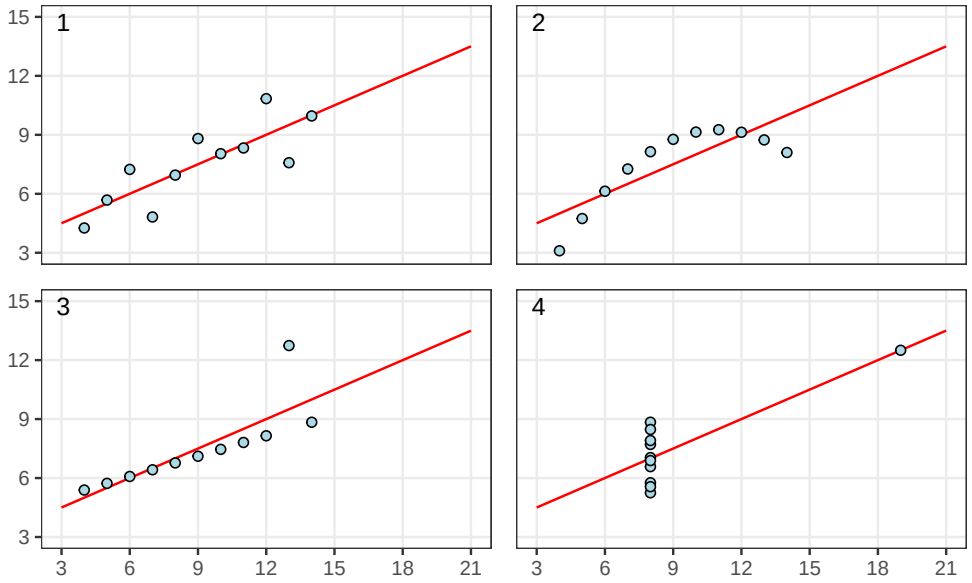
$$\hat{\alpha} = 58.9 \quad \text{und} \quad \hat{\beta} = 0.875$$

der unten dargestellten Ausgleichsgeraden.



Anscombes Quartett

Einen Datensatz durch eine Ausgleichsgerade anzunähern ist natürlich nur dann sinnvoll, wenn für die betrachteten Werte ein linearer Zusammenhang zu Grunde liegt. Darauf, dass es immer wichtig ist, neben der Ausgleichsgeraden auch die zugrunde liegenden Daten zu betrachten, hat der Statistiker John Anscombe im Jahr 1973 hingewiesen. Er hat hierzu die unten dargestellten vier Datensätze verwendet, die alle zur selben Regressionsgeraden mit $\hat{\alpha} = 3$ und $\hat{\beta} = 1/2$ führen.



Zu den vier Datensätzen halten wir fest:

1. Ausgleichsgerade sinnvoll, offenbar liegt ein linearer Zusammenhang vor.
2. Der Zusammenhang ist quadratisch, Ausgleichsgerade sinnlos.
3. Die Ausgleichsgerade wird durch einen Ausreißer verfälscht.
4. Ohne den Ausreißer könnte keine Ausgleichsgerade bestimmt werden.

Bestimmtheitsmaß

Um die Güte des linearen Modells zu beurteilen, untersucht man, wie die Werte auf der Ausgleichsgeraden $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ und die ursprünglichen Werte y_i um den Mittelwert $\bar{y}$ herum streuen. Hierzu werden die Differenzen zum Mittelwert quadriert und aufsummiert. Der Quotient aus den beiden Quadratsummen ist das so genannte Bestimmtheitsmaß. Je näher die Punkte an der Ausgleichsgeraden liegen, umso mehr stimmen die beiden Differenzen überein. Liegen alle Punkte exakt auf der Ausgleichsgeraden, dann ist das Bestimmtheitsmaß gleich eins.

Definition 0.4 (Bestimmtheitsmaß)**.** Für die Werte $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ erfasst das Bestimmtheitsmaß

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

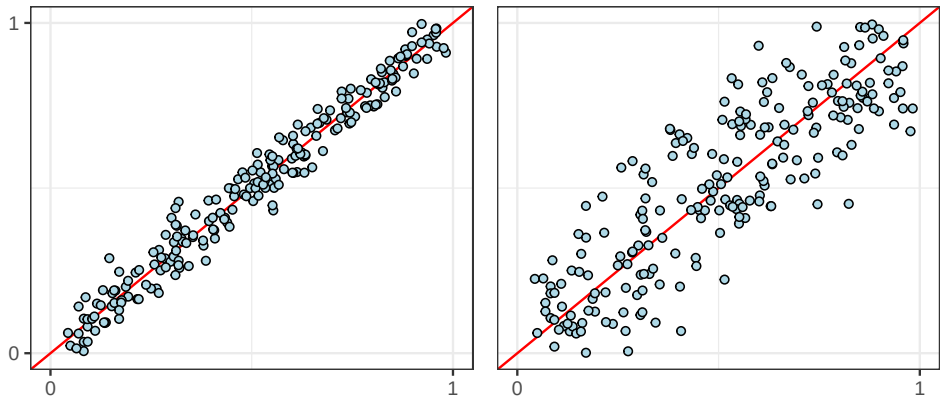
mit $0 \leq R^2 \leq 1$ die Güte des linearen Modells. Dabei ist

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i.$$

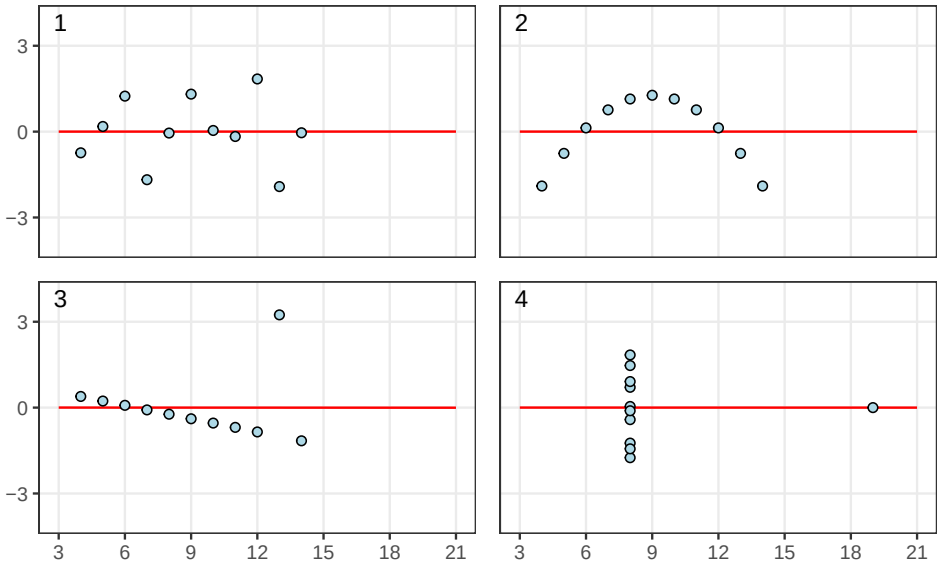
Es lässt sich zeigen, dass es sich bei dem Bestimmtheitsmaß um das Quadrat des Korrelationskoeffizienten r handelt, es gilt also $R = r^2$.

Damit kann der Korrelationskoeffizient auch in Bezug auf das lineare Regressionsmodell interpretiert werden: Der quadrierte Korrelationskoeffizient entspricht dem Anteil der Gesamtstreuung, der durch das lineare Modell erklärt werden kann.

Als Beispiel betrachten wir die beiden nachfolgend dargestellten Datensätze. Wir erhalten die Bestimmtheitsmaße links mit $R^2 = 0.97$ und rechts mit $R^2 = 0.73$. Die unterschiedliche Variabilität um die Ausgleichsgerade herum wird also sehr gut erfasst.



Gegenbeispiel: Für die Datensätze aus Anscombes Quartett ergeben sich in allen vier Fällen für das Bestimmtheitsmaß Werte von $R^2 = 0.67$, so dass wir hier keinerlei Aufschluss darüber erhalten, ob das lineare Modell sinnvoll ist oder nicht. Im Zweifelsfall sollte man daher immer die Residuen ε_i plotten, wie das hier für das Anscombe-Quartett dargestellt ist.



Anmerkung: Der Wertebereich für den Korrelationskoeffizienten r und das Bestimmtheitsmaß R^2 ist jeweils das Intervall $[0, 1]$. Es liegt daher nahe, die Werte als prozentuale Größen zu interpretieren. Man spricht dann zum Beispiel davon, dass eine Korrelation von 40 % besteht, oder dass eine Variabilität zu 16 % durch das lineare Modell erklärt werden kann.

Ausblick: Nichtlineare und lokale Regression

Liegt den Daten ein nichtlinearer Zusammenhang zu Grunde, dann kann sinnvoll sein, mit nichtlinearen Funktionen zu arbeiten. Zum Beispiel kann ein quadratisches Modell

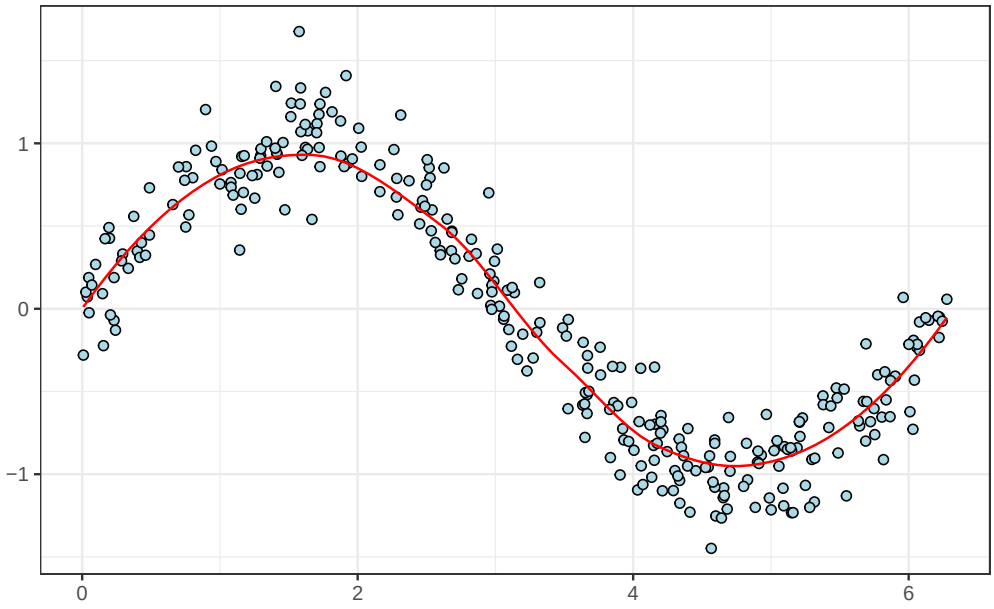
$$Y = \alpha + \beta X + \gamma X^2 + \varepsilon$$

oder ein exponentieller Ansatz

$$Y = \alpha \exp(\beta X) + \varepsilon$$

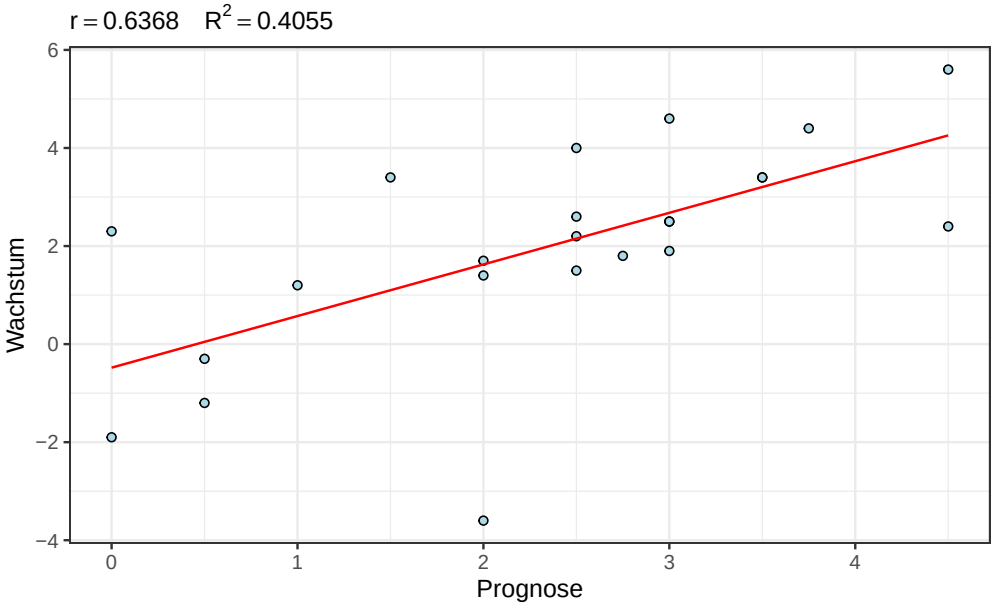
verwendet werden. Im quadratischen Fall bestimmen sich die Koeffizienten α, β, γ aus einem linearen Gleichungssystem mit drei unbekannten. Für das exponentielle Modell ist ein nichtlineares Gleichungssystem zu lösen.

Alternativ kann mit der LOESS-Methode eine Regressionskurve berechnet werden, die auch komplexere Zusammenhänge zwischen Merkmalen erfassen kann. Das Akronym LOESS steht dabei für *Locally Estimated Scatterplot Smoothing*. Unten dargestellt ist ein Datensatz mit Werten, die um eine Sinusfunktion herum streuen. Die zugehörige Regressionskurve bildet den zugrunde liegenden Zusammenhang sehr gut ab.



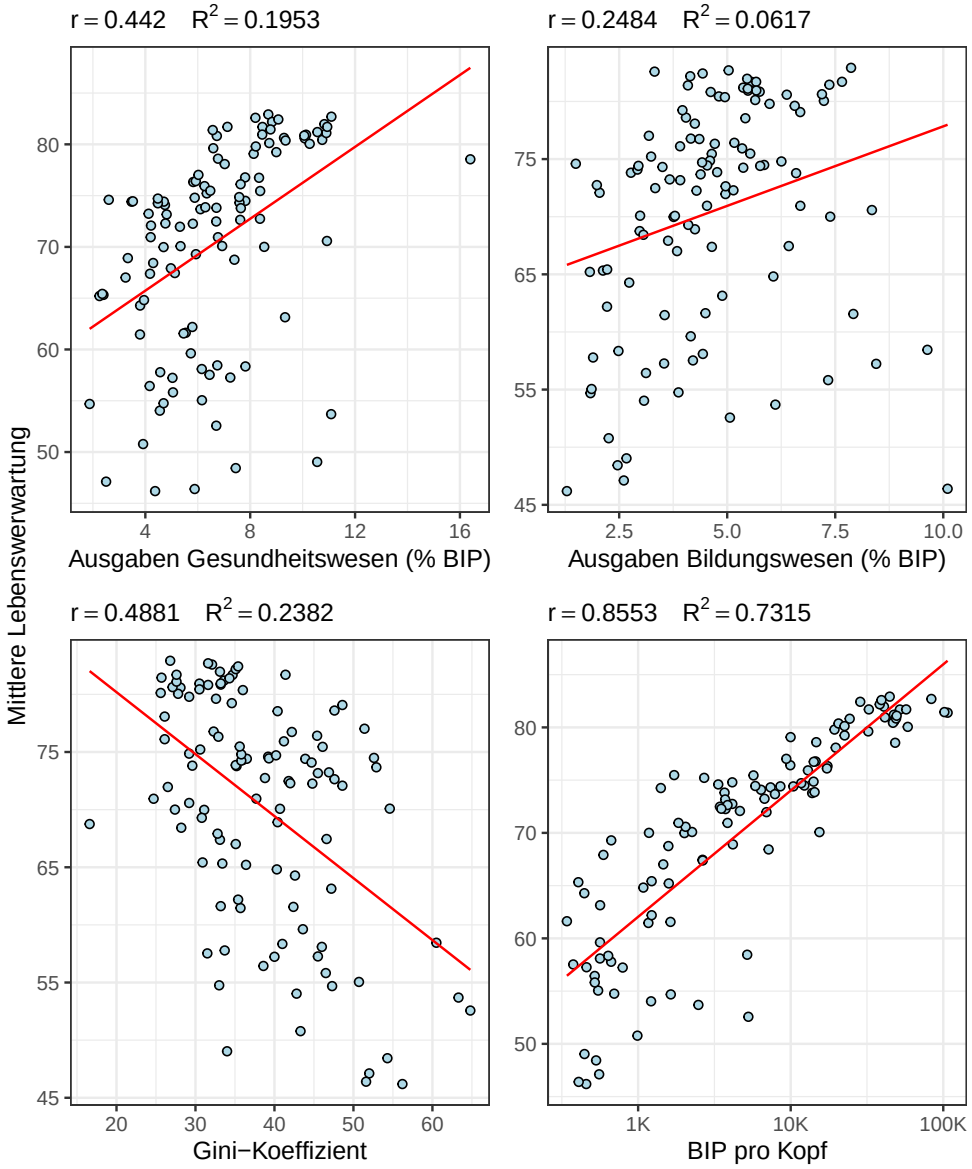
Beispiele

Abschließend werden die Datensätze zu den Prognosen des Sachverständigenrats und der Lebenserwartung mit den oben diskutierten Methoden untersucht.



Für die Prognose des Sachverständigenrats besteht mit 64 % eine mittlere Korrelation mit dem realen Wirtschaftswachstum. Entsprechend lassen sich 40 % der Variabilität des Wachstums durch den unterstellten linearen Zusammenhang erklären.

In den nachfolgend dargestellten Auswertungen zur mittleren Lebenserwartung auf Grundlage der Weltbankdaten zeigt sich, abgesehen vom BIP pro Kopf, für alle Indikatoren eine schwache Korrelation. Dabei sind die Ausgaben für das Gesundheitswesen und der Gini-Koeffizient etwa gleich stark mit der Lebenserwartung korreliert, der Zusammenhang zwischen den Ausgaben für Bildung ist erwartungsgemäß schwächer. Demgegenüber zeigt sich für die BIP pro Kopf (in einer logarithmischen Skalierung) mit 86 % eine starke Korrelation mit der Lebenserwartung.



Fahrmeir, Ludwig, Christian Heumann, Rita Künstler, Iris Pigeot, und Gerhard Tutz. 2016. *Statistik, Der Weg zur Datenanalyse*. 8. Auflage. Springer Spektrum.

Index