9. Korrelation und Regression

Inhalt

In diesem Kapitel wenden wir uns wieder der Beschreibenden Statistik zu, um die Analyse ***multivariater Daten*** zu behandeln. Liegen multivariate Daten vor, also Stichproben mit mehreren gleichzeitig gemessenen Merkmalen, so ist es zunächst natürlich möglich, jedes Merkmal einzeln mit den Methoden für univariate Daten zu behandeln. Interessanter ist es jedoch meistens, Zusammenhänge und Abhängigkeiten zwischen verschiedenen Merkmalen zu untersuchen; dazu dienen Korrelations- und Regressionsverfahren. In der Vorlesung werden wir uns auf ***bivariate Daten*** beschränken.

Bivariate Daten Stichproben mit zwei gleichzeitig gemessenen Merkmalen, z. B. in Form von Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$

Korrelation Erkennen von Zusammenhängen bzw. Abhängigkeiten zwischen verschiedenen Merkmalen einer Stichprobe mit multivariaten Daten, für bivariate Daten z. B. zwischen x und y

Regression Bestimmen von ***Modellen*** zur Beschreibung solcher Zusammenhänge, für bivariate Daten z. B. in Form von Funktionen $x \mapsto y(x)$

9.1. Zweidimensionale Häufigkeitsverteilungen

9.1.1. Streudiagramm

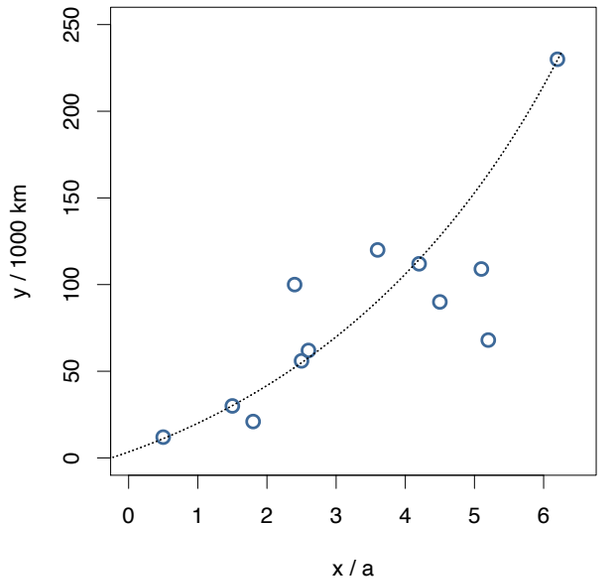
Eine Urliste bivariater Daten kann in Form eines ***Streudiagramms*** graphisch dargestellt werden.

Beispiel 9.1 (Fuhrpark)
Für zwölf Kraftfahrzeuge eines Fuhrparks werden die Merkmale x: „Alter in Jahren“ und y: „Tachometerstand in 1000 km“ erfasst.

i	x/a	y/1000 km	i	x/a	y/1000 km
1	1.5	30	7	1.8	21
2	5.2	68	8	4.2	112
3	4.5	90	9	6.2	230
4	0.5	12	10	3.6	120
5	2.4	100	11	2.5	56
6	2.6	62	12	5.1	109

Quelle: Sachs 2018

Graphische Darstellung der bivariaten Daten (x_i, y_i) in einem ***Streudiagramm***:



9.1.2. Zweidimensionale absolute Häufigkeit

Für die folgenden Betrachtungen ist es nützlich, die Definitionen der Häufigkeitsverteilungen auf zwei (oder mehr) Dimensionen zu erweitern.

Definition 9.2 (Zweidimensionale absolute Häufigkeit)
Seien X, Y Merkmale der statistischen Elemente einer Stichprobe vom Umfang n und seien $(x_1, y_1), \ldots, (x_n, y_n)$, die Wertepaare der Merkmalsausprägungen.

Seien $m \leq n$ bzw. $\ell \leq n$ die jeweilige Anzahl unterschiedlicher Ausprägungen des Merkmals X bzw. Y und seien $a_1, \ldots, a_m$ bzw. $b_1, \ldots, b_\ell$ die jeweiligen (geordneten) unterschiedlichen Ausprägungen.

Die ***zweidimensionale absolute Häufigkeit*** h_{jk} des Ausprägungspaars (a_j, b_k) ist die Anzahl statistischer Elemente mit $(x_i, y_i) = (a_j, b_k)$,

$$h_{jk} = |\{i : (x_i, y_i) = (a_j, b_k),\ 1 \leq i \leq n\}|, \quad j = 1, \ldots, m, \quad k = 1, \ldots, \ell.$$

Analog dazu lassen sich zweidimensionale relative, klassierte und kumulierte Häufigkeiten definieren. Ebenso lassen sich höherdimensionale Häufigkeitsverteilungen definieren.

Definition 9.3 (Randhäufigkeiten)
Seien die gleichen Voraussetzungen gegeben wie in Definition 9.2, und sei h_{jk} die zugehörige zweidimensionale absolute Häufigkeit.

Dann sind die ***Randhäufigkeiten*** $h_{j\cdot}$ bzw. $h_{\cdot k}$ definiert als die jeweilige Anzahl statistischer Elemente mit $x_i = a_j$ bzw. $y_i = b_k$,

$$h_{j\cdot} = |\{i : x_i = a_j,\ 1 \le i \le n\}| = \sum_{k=1}^{\ell} h_{jk}, \quad j = 1, \ldots, m,$$
$$h_{\cdot k} = |\{i : y_i = b_k,\ 1 \le i \le n\}| = \sum_{j=1}^{m} h_{jk}, \quad k = 1, \ldots, \ell.$$

9.1.3. Kontingenztafel

Zweidimensionale Häufigkeiten und Randhäufigkeiten lassen sich in einer ***Kontingenztafel*** (auch ***zweidimensionale Häufigkeitstabelle*** darstellen. Wie in einer Matrix bestimmen der erste Index die Zeile und der zweite Index die Spalte.

		Y				
		b_1	b_2	...	b_ℓ	$h_{j\cdot}$
	a_1	h_{11}	h_{12}	...	$h_{1\ell}$	$h_{1\cdot}$
X	a_2	h_{21}	h_{22}	...	$h_{2\ell}$	$h_{2\cdot}$
	...	...	...	...	...	...
	a_m	h_{m1}	h_{m2}	...	$h_{m\ell}$	$h_{m\cdot}$
	$h_{\cdot k}$	$h_{\cdot 1}$	$h_{\cdot 2}$	...	$h_{\cdot \ell}$	n

9.2. Kovarianz und Korrelation

9.2.1. Empirische Kovarianz

Die empirische Kovarianz ist eine nichtstandardisierte Maßzahl für den monotonen (oder linearen) Zusammenhang zweier Merkmale.

Definition 9.4 (Empirische Kovarianz)
Gegeben sei eine Stichprobe vom Umfang n mit den Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$.

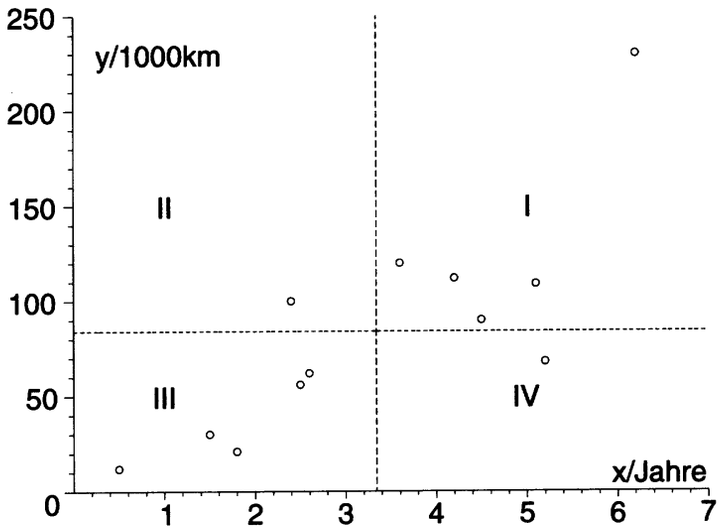
Dann ist die ***empirische Kovarianz*** s_{xy} definiert als

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

Dabei bezeichnen $\bar{x}$, $\bar{y}$ die arithmetischen Mittelwerte von $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$.

Die Kovarianz kann in gewisser Hinsicht als Verallgemeinerung der Varianz angesehen werden: Für Wertepaare der Form (x_i, x_i) (also mit $y_i = x_i$ für $1 \leq i \leq n$) entspricht die Kovarianz s_{xx} der Varianz s_x^2.

Um die Bedeutung der Kovarianz zu verstehen, teilt man das Koordinatensystem mit Hilfe von zwei Geraden durch den ***Schwerpunkt*** $(\bar{x}, \bar{y})$ in vier Quadranten ein.



Bildquelle: Sachs 2018

Für Punkte (x_i, y_i) in den Quadranten I bis IV gilt:

I III	$x_i > \bar{x} \wedge y_i > \bar{y}$ $x_i < \bar{x} \wedge y_i < \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y}) > 0$
II IV	$x_i < \bar{x} \wedge y_i > \bar{y}$ $x_i > \bar{x} \wedge y_i < \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y}) < 0$

Demnach folgt:

- $s_{xy} > 0$: Es überwiegen Punkte in den Quadranten I und III, die Punkte verlaufen von links unten nach rechts oben.
- $s_{xy} < 0$: Es überwiegen Punkte in den Quadranten II und IV, die Punkte verlaufen von links oben nach rechts unten.

Beispiel : Fuhrpark

Für die Daten aus Beispiel 9.1 erhalten wir $s_{xy} \approx 81.6$.

Satz 9.5 (Rechenregeln für die empirische Kovarianz)
Für die empirische Kovarianz gelten die Rechenregeln

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right),$$
$$s_{xy} = \frac{1}{n-1}\left(\sum_{j=1}^{m}\sum_{k=1}^{\ell} a_j b_k h_{jk} - n\overline{xy}\right).$$

Satz 9.6 (Perfekte lineare Abhängigkeit)
Für die empirische Kovarianz gilt

$$\left|s_{xy}\right| \leq s_x s_y.$$

Gleichheit gilt genau dann, wenn zwischen (x_i) und (y_i) eine ***perfekte lineare Abhängigkeit*** besteht,

$$\left|s_{xy}\right| = s_x s_y \quad \Leftrightarrow \quad y_i = a + b x_i, \quad 1 \leq i \leq n$$

mit festen Parametern a und b.

9.2.2. Linearer Korrelationskoeffizient

Die Kovarianz hängt von der Maßeinheit der Merkmale ab, was ihre absoluten Werte schwer interpretierbar macht. Durch Normierung mit Hilfe der Standardabweichungen gelangt man zu einem maßstabsunabhängigen Korrelationskoeffizienten mit Werten zwischen -1 und $+1$.

Definition 9.7 (Linearer Korrelationskoeffizient)
Gegeben sei eine Stichprobe vom Umfang n mit den Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$.

Dann ist der ***lineare Korrelationskoeffizient*** (oder ***Pearsonsche Korrelationskoeffizient***) r_{xy} definiert als

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Dabei bezeichnen s_{xy} die empirische Kovarianz, s_x, s_y die empirischen Standardabweichungen und $\overline{x}, \overline{y}$ die arithmetischen Mittelwerte.

Satz 9.8 (Perfekte lineare Abhängigkeit)
Für den linearen Korrelationskoeffizienten gilt

$$-1 \leq r_{xy} \leq +1.$$

Gleichheit gilt jeweils genau dann, wenn zwischen (x_i) und (y_i) eine ***perfekte lineare Abhängigkeit*** besteht,

$$r_{xy} = -1 \quad \Leftrightarrow \quad y_i = a + b x_i, \quad 1 \le i \le n \quad \text{mit} \quad b < 0,$$
$$r_{xy} = +1 \quad \Leftrightarrow \quad y_i = a + b x_i, \quad 1 \le i \le n \quad \text{mit} \quad b > 0,$$

mit festen Parametern a und b.

Der Korrelationskoeffizient liefert eine Aussage über die Abhängigkeit zwischen zwei Merkmalen X und Y:

- $r_{xy} \approx +1$: X und Y sind *stark positiv korreliert*.
- $r_{xy} \approx -1$: X und Y sind *stark negativ korreliert*.
- $r_{xy} \approx 0$: X und Y sind *unkorreliert*.

Genaue Grenzen anzugeben ist schwierig; ein Korrelationskoeffizient mit $|r_{xy}| < 0.7$ deutet jedoch kaum noch auf einen linearen Zusammenhang hin, wie die folgenden Beispiele zeigen. Wichtig ist auf jeden Fall die folgende Überlegung.

Ein betragsmäßig hoher Korrelationskoeffizient erlaubt allein noch *keine* Aussage über einen kausalen Zusammenhang (im Sinne von Ursache und Wirkung) zwischen X und Y:

- X kann Ursache von Y sein.
- Y kann Ursache von X sein.
- X und Y können gemeinsam von einer unbekannten Ursache Z (oder von weiteren Ursachen) abhängen.
- X und Y können sich zufällig ähnlich verhalten.

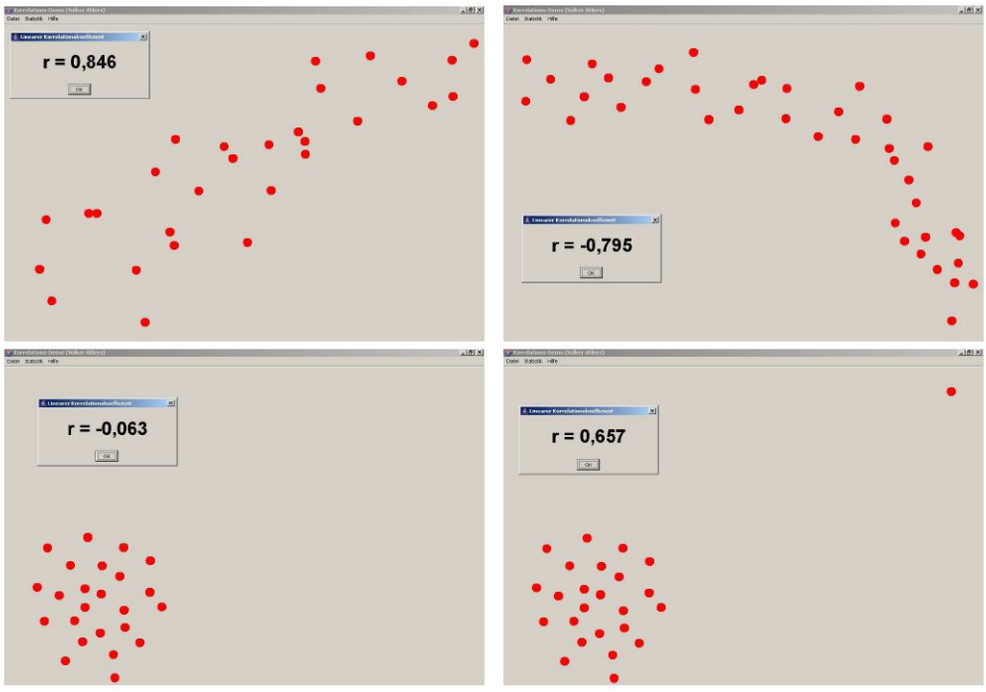
Beispiel 9.9 (Fuhrpark)
Für die Daten aus Beispiel 9.1 erhalten wir

$$s_{xy} \approx 81.6, \quad s_x \approx 1.73, \quad s_y \approx 58.7$$

und somit

$$r_{xy} \approx 0.804.$$

Beispiel 9.10
Die folgende Abbildung gibt weitere Beispiele an und weist zugleich auf einige Probleme hin.



Mögliche Probleme des linearen Korrelationskoeffizienten:

- r_{xy} kann auch für nichtlineare, aber streng monotone Zusammenhänge einen betragsmäßig hohen Wert haben (Abb. rechts oben). Dies kann man als Vorteil oder als Nachteil ansehen.
- r_{xy} wird sehr stark von Ausreißern bestimmt (Abb. rechts unten).
- Für ***stückweise lineare Abhängigkeiten*** kann r_{xy} einen betragsmäßig kleinen Wert haben, z. B. gilt

$$y_i = |x_i| \quad \wedge \quad x_i \text{ gleichmäßig verteilt in } [-1,1] \quad \Rightarrow \quad r_{xy} = 0.$$

Fazit: Der lineare Korrelationskoeffizient ist nur für lineare oder zumindest deutlich streng monotone Zusammenhänge geeignet.

9.2.3. Rang-Korrelationskoeffizient

Eine robuste Alternative zum linearen Korrelationskoeffizienten stellt der ***Rang-Korrelationskoeffizient*** dar.

Definition 9.11 (Rang-Korrelationskoeffizient)
Gegeben sei eine Stichprobe vom Umfang n mit den Wertepaaren $((x_1,y_1),\ldots,(x_n,y_n))$.

Es sei $R_{x,i}$ der ***Rang*** von x_i innerhalb der Werte $x_1,\ldots,x_n$, d. h. seine Position, wenn alle Werte x_j aufsteigend der Größe nach geordnet wurden. Ebenso sei $R_{y,i}$ der Rang von y_i innerhalb der Werte $y_1,\ldots,y_n$.

- soc hier aufsteigend der Größe

Rx: index

Dann ist der (Spearmansche) ***Rang-Korrelationskoeffizient*** $r_{s,xy}$ definiert als

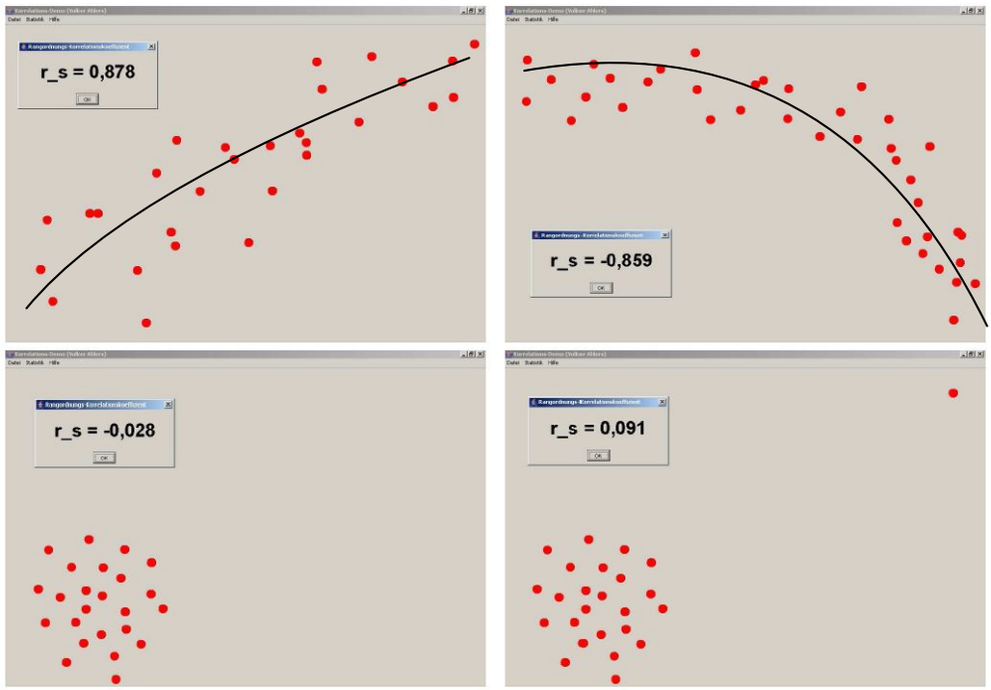
$$r_{s,xy} = r_{R_x R_y} = \frac{\sum_{i=1}^{n}(R_{x,i} - \overline{R}_x)(R_{y,i} - \overline{R}_y)}{\sqrt{\sum_{i=1}^{n}(R_{x,i} - \overline{R}_x)^2}\sqrt{\sum_{i=1}^{n}(R_{y,i} - \overline{R}_y)^2}}.$$

Dabei bezeichnet $r_{R_x R_y}$ den linearen Korrelationskoeffizienten der Rangpaare $(R_{x,i}, R_{y,i})$.

Beispiel zur Erläuterung des Rangs:

$$(x_i) = (72, 34, 19, 51, 80) \quad \Rightarrow \quad (R_{x,i}) = (4, 2, 1, 3, 5).$$

Beispiel 9.12
Die folgende Abbildung zeigt die Rang-Korrelationskoeffizienten für die Daten aus Beispiel 9.10.



Eigenschaften des Rang-Korrelationskoeffizienten:

- $r_{s,xy}$ hat auch für nichtlineare, aber monotone Zusammenhänge einen betragsmäßig hohen Wert (Abb. rechts oben). Er stellt eine ***nicht-parametrische*** (auch ***parameterfreie***) Methode dar, d. h. ihm liegt keine feste Modellstruktur zugrunde.
- $r_{s,xy}$ ist robust gegenüber Ausreißern (Abb. rechts unten).
- Für ***stückweise lineare Abhängigkeiten*** (z. B. $y_i = |x_i|$) kann auch $r_{s,xy}$ einen betragsmäßig kleinen Wert haben.

Fazit: Der Rang-Korrelationskoeffizient ist für beliebige streng monotone Zusammenhänge geeignet.

9.3. Regression

9.3.1. Modellierung

Regressionsverfahren (von lateinisch *regredi*: zurückgehen) verwenden Modelle, um Zusammenhänge zwischen Daten zu beschreiben (siehe Abschnitt 1.2.3). Eine zentrale Aufgabe besteht dabei in der Bestimmung der Modellparameter.

Mathematische Modelle können u. a. linear, quadratisch oder exponentiell sein.

- Lineares Modell: $y = a + bx$, z. B. $U(I) = RI$
- Quadratisches Modell: $y = a + bx + cx^2$, z. B. $s(t) = s_0 + vt + \frac{1}{2}at^2$
- Exponentielles Modell: $y = ae^{bx}$, z. B. $u(t) = U_0 e^{-t/(RC)}$

Wir beschränken uns zunächst auf lineare Modelle.

Gegeben seien Wertepaare (x_i, y_i), die einen linearen Zusammenhang zwischen x und y vermuten lassen.

Der passende lineare Modellansatz besteht aus der Geradengleichung

$$y(x) = a + bx,$$

in der die Parameter a und b (also y-Achsenabschnitt und Steigung) so zu bestimmen sind, dass die Abweichungen

$$|y_i - y(x_i)| = |y_i - a - bx_i|$$

möglichst klein sind – in einem noch genauer zu definierenden Sinn.

9.3.2. Lineare Regression

Die Schätzung der Modellparameter erfolgt mit Hilfe der ***Methode der kleinsten Quadrate*** (engl. *least squares method*, A.-M. LEGENDRE 1806, C. F. GAUSS 1809). GAUSS gelang es 1801 unter Verwendung dieser Methode, die durch die Sonne verdeckte Bahn des Asteroiden Ceres mit hoher Genauigkeit vorherzusagen. Später konnte er zeigen, dass dieser so genannte *Kleinste-Quadrate-Schätzer* für lineare Modelle unter bestimmten, häufig erfüllten Voraussetzungen optimal ist (Satz von Gauß-Markov).

Bestimme die Parameter a und b so, dass die Summe der Fehlerquadrate

$$q(a,b) = \sum_{i=1}^{n} (y_i - y(x_i))^2$$

minimal wird.

Wir geben zunächst das Ergebnis in einem Satz an und zeigen im Beweis, wie dieses mit Hilfe der partiellen Ableitungen nach a und b bestimmt wird.

Satz 9.13 (Lineare Regression)
Gegeben sei eine Stichprobe vom Umfang n mit den Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$.

Dann ist die ***Regressionsgerade*** $y(x) = a + bx$, für die

$$q(a,b) = \sum_{i=1}^{n} (y_i - y(x_i))^2$$

minimal wird, gegeben durch die ***linearen Regressionsparameter***

$$\hat{b} = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Dabei bezeichnen $\bar{x}$, $\bar{y}$ die arithmetischen Mittelwerte, s_x^2 die empirische Varianz, s_{xy} die empirische Kovarianz und r_{xy} den linearen Korrelationskoeffizienten.

Die Schreibweise $\hat{a}$, $\hat{b}$ („a-Dach“, „b-Dach“) deutet an, dass es sich bei diesen Werten um ***Schätzer*** handelt. Schätzer werden wir in der Schließenden Statistik im nächsten Kapitel genauer behandeln.

Beweis: Für den folgenden Beweis sind Kenntnisse der Analysis erforderlich, die erst in der Lehrveranstaltung *Mathematik 3* vermittelt werden. Er ist daher nur der Vollständigkeit halber für Interessierte angegeben.

Um das Minimum der Summe der Fehlerquadrate, also der Funktion

$$q(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

zu bestimmen, setzen wir die partiellen Ableitungen $\partial q/\partial a$ und $\partial q/\partial b$ gleich null und lösen das resultierende lineare Gleichungssystem für a und b. Zunächst erhalten wir

$$\begin{aligned}
\frac{\partial q(a,b)}{\partial a} &= \sum_{i=1}^{n} (-1) \cdot 2(y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b \sum_{i=1}^{n} x_i \right) \\
&= -2 \left(\sum_{i=1}^{n} y_i - na - b \sum_{i=1}^{n} x_i \right)
\end{aligned}$$

und damit

$$\frac{\partial q(a,b)}{\partial a} = 0 \quad \Longleftrightarrow \quad a = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{b}{n} \sum_{i=1}^{n} x_i = \bar{y} - b\bar{x}.$$

Diesen Ausdruck verwenden wir zur Berechnung von b. Es ist

$$\begin{aligned}\frac{\partial q(a,b)}{\partial b} &= \sum_{i=1}^{n}(-x_i)\cdot 2(y_i - a - bx_i)\\ &= -2\left(\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2\right)\\ &= -2\left(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + b\bar{x}\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2\right)\\ &= -2\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y} + b\left(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2\right)\right)\end{aligned}$$

und somit

$$\frac{\partial q(a,b)}{\partial b} = 0 \quad\Longleftrightarrow\quad b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{(n-1)s_{xy}}{(n-1)s_x^2} = \frac{s_{xy}}{s_x^2}.$$

Schließlich folgt mit der Definition des linearen Korrelationskoeffizienten r_{xy}

$$b = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}\,s_y}{s_x^2\,s_y} = r_{xy}\frac{s_y}{s_x}.$$

■

Beispiel : Fuhrpark

Für die Daten aus Beispiel 9.1 und 9.9 erhalten wir

$$s_{xy} \approx 81.6, \qquad s_x^2 \approx 2.99, \qquad \bar{x} \approx 3.34, \qquad \bar{y} \approx 84.1$$

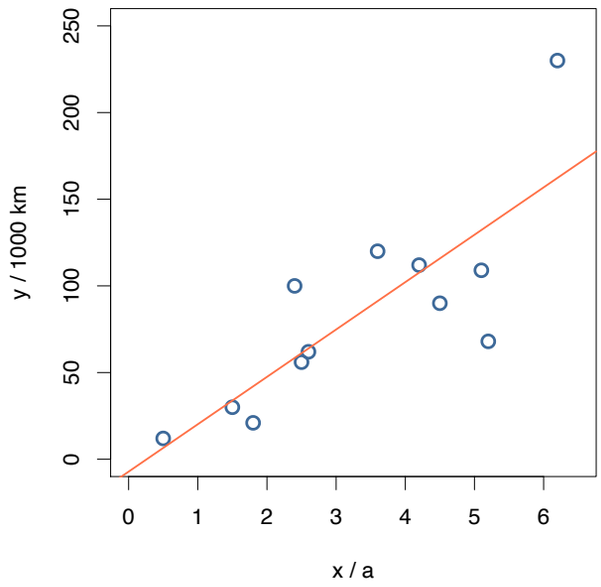
und somit

$$\hat{b} \approx 27.3, \qquad \hat{a} \approx -7.17,$$

also die Regressionsgerade

$$y(x) \approx -7.17 + 27.3x.$$

Graphische Darstellung der Regressionsgerade zusammen mit dem Streudiagramm aus Beispiel 9.1:



Die Regressionsgerade kann auch in der Form

$$y - \bar{y} = \hat{b}(x - \bar{x})$$

geschrieben werden. Die Gerade verläuft also durch den ***Schwerpunkt*** $(\bar{x}, \bar{y})$ der Stichprobe.

Definition 9.14 (Residuum)
Gegeben seien eine Stichprobe vom Umfang n mit den Wertepaaren $(x_1, y_1), \ldots, (x_n, y_n)$ sowie lineare Regressionsparameter $\hat{a}$, $\hat{b}$.

Dann ist das ***i-te Residuum*** r_i definiert als die Differenz

$$r_i = y_i - \hat{y}_i \quad \text{mit} \quad \hat{y}_i = \hat{a} + \hat{b}\, x_i.$$

Beispiel : Anscombe's Quartet

Die folgenden vier Datensätze der Form (x_i, y_i) stammen von F. J. ANSCOMBE und sind bekannt als *Anscombe's quartet*.

Sind die Datensätze statistisch vergleichbar?

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Quelle: Tufte, E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

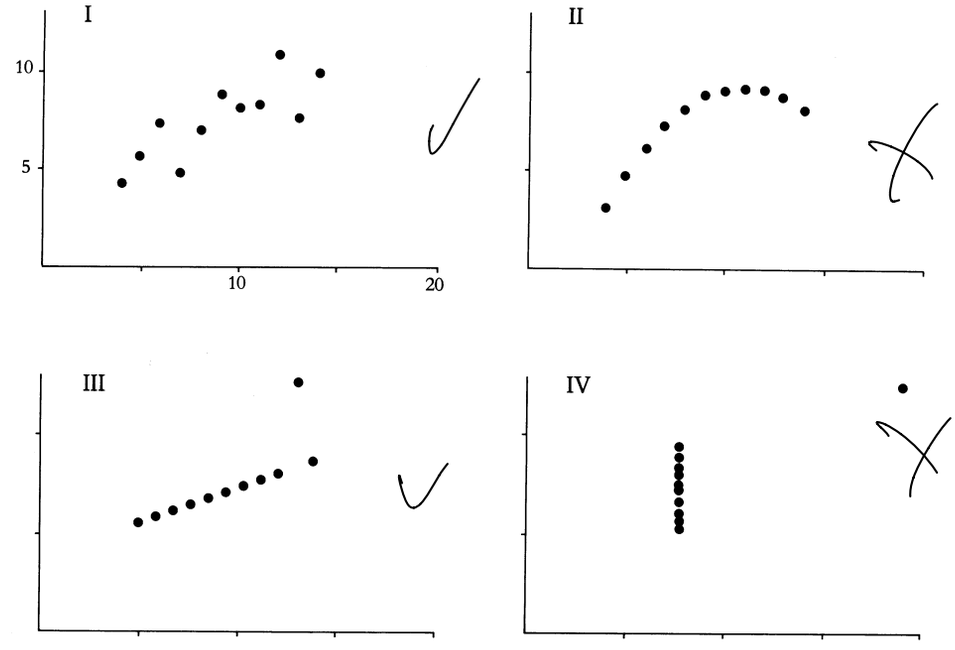
Für alle vier Datensätze gilt

$$n = 11, \qquad \bar{x} = 9.0, \qquad \bar{y} \approx 7.5.$$

Eine lineare Korrelations- und Regressionsanalyse liefert für alle vier Datensätze

$$r_{xy} \approx 0.82, \qquad \hat{a} \approx 3.0, \qquad \hat{b} \approx 0.5, \qquad q(\hat{a}, \hat{b}) \approx 13.75.$$

Die graphische Darstellung zeigt jedoch wesentliche Unterschiede.



Bildquelle: Tufte, E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

✓ passt

✗ passt. nicht

Fazit: Das der linearen Korrelations- und Regressionsanalyse zugrunde liegende Modell ist für die Datensätze II und IV ungeeignet, obwohl der lineare Korrelationskoeffizient $r_{xy} = 0.82$ auf eine lineare Korrelation hinzudeuten scheint. Außerdem wird deutlich, dass der Ausreißer in Datensatz III starken Einfluss auf die Ergebnisse der Korrelations- und Regressionsanalyse hat: Ohne Berücksichtigung dieses Ausreißers liegen die Datenpunkte fast exakt auf einer Geraden ($r_{xy} \approx 1$, $q(\hat{a},\hat{b}) \approx 0$, deutlich andere Werte der Regressionsparameter $\hat{a} \approx 3.95$ und $\hat{b} \approx 0.35$ als mit Ausreißer).

9.3.3. Allgemeine lineare Regression und nichtlineare Regression

Die Regressionsanalyse lässt sich auch auf Polynome höheren Grades anwenden, z. B. quadratische Modelle

$$y(x) = a_0 + a_1 x + a_2 x^2.$$

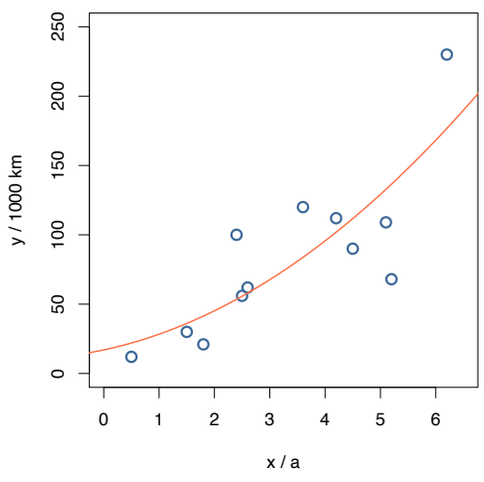
Da die Modellparameter a_i linear in das Modell eingehen, handelt es sich um eine ***allgemeine lineare Regression***. Die Methode der kleinsten Quadrate ist hierbei äquivalent zu ***linearen Ausgleichsproblemen*** mit überbestimmten linearen Gleichungssystemen (mehr Gleichungen als unbekannte Modellparameter), die sich z. B. mit Hilfe der ***QR-Zerlegung*** oder der ***Singulärwertzerlegung*** lösen lassen.

Beispiel : Fuhrpark

Graphische Darstellung der quadratischen Regressionskurve

$$y(x) \approx 17.0 + 8.55\,x + 2.78\,x^2$$

für die Daten aus Beispiel 9.1:



Weiterhin lässt sich die lineare Regression auf multivariate Daten $(x_1,\ldots,x_m)$ ausweiten, indem Modelle der Form

$$x_m(x_1,\ldots,x_{m-1}) = b_0 + b_1 x_1 + \ldots + b_{m-1} x_{m-1}$$

verwendet werden. Auch hierbei sind lineare Ausgleichsprobleme zu lösen.

Schließlich lassen sich mit der Methode der kleinsten Quadrate auch ***nichtlineare Modellansätze*** behandeln (in welche die Modellparameter nichtlinear eingehen), z. B. exponentielle Modelle

$$y(x) = a\,e^{bx}.$$

Die Modellparameter werden in diesem Fall i. a. iterativ bestimmt.

9.4. Korrelation und Regression in R

- `cov(x, y)`: empirische Kovarianz s_{xy}
- `cor(x, y, method=...)`:
  - `method="pearson"` (Voreinstellung): linearer Korrelationskoeffizient r_{xy}
  - `method="spearman"`: Rang-Korrelationskoeefizient $r_{s,xy}$
- `lm(y ~ x)` (*linear model*): Lineare Regression $y(x) = a + bx$

```
x <- c(1,2,3,5)
y <- c(7,5,3,2)
r.xy <- cor(x, y)
rs.xy <- cor(x, y, method="spearman")

linReg <- lm(y ~ x)
plot(x, y)
abline(linReg)
```

Erläuterung:

- `y ~ x` ist ein Objekt der R-Klasse `formula` zur Beschreibung des Modells, hier einer linearen Abhängigkeit $y(x)$. Ein skalarer Term wird automatisch ergänzt, wir hätten gleichbedeutend `y ~ 1 + x` schreiben können.
- Die Funktion `abline()` nutzt das lineare Modell, um eine Gerade zu zeichnen.

Die Funktion `lm()` unterstützt auch allgemeine und multivariate lineare Regression. Im folgenden Beispiel wird das quadratische Modell

$$y(x) = a_0 + a_1 x + a_2 x^2$$

verwendet.

```
linReg2 <- lm(y ~ x + I(x^2))
xSeq = seq(1, 5, length=100)
lines(xSeq, predict(linReg2, data.frame(x=xSeq)))
```

Erläuterung:

- In `y ~ x + I(x^2)` ist die Funktion `I()` (*identity*) erforderlich, weil der Operator `^` in Objekten der Klasse `formula` eine andere Bedeutung als die Potenzierung hat. Einzelheiten sind der R-Literatur zu entnehmen.
- Die Funktion `predict()` nutzt das lineare Modell zur Berechnung von Zwischenwerten im Intervall $[1,5]$.