3. Mehrdimensionale (multivariate) Datenanalyse

Andreas Eberle
Institut für angewandte Mathematik

Oktober 2008

- Wir wollen nun den Zusammenhang von zwei Merkmalen X und Y mit möglichen Merkmalsausprägungen a_k bzw. b_l untersuchen.
- Dazu genügt es nicht, die (relativen) Häufigkeiten der verschiedenen Merkmalsausprägungen für X und Y zu kennen, sondern wir müssen wissen, welche Werte X annimmt, wenn Y eine bestimmte Merkmalsausprägung hat, bzw. umgekehrt.
- Wir müssen also die kombinierten Merkmalsausprägungen

$$(x_i, y_i) = (X(\omega_i), Y(\omega_i))$$

kennen, oder zumindest wissen, wie oft diese die möglichen Ausprägungen (a_k, b_l) annehmen.

3.1. Kontingenztabelle

Absolute und relative Häufigkeiten

- Die *Häufigkeit* der kombinierten Merkmalsausprägung (a_k, b_l) ist

  n_{kl} = Anzahl der stat. Einheiten ω_i mit $x_i = a_k$ und $y_i = b_l$.

- Die *relative Häufigkeit* von (a_k, b_l) ist

$$h_{kl} = \frac{n_{kl}}{n} = \frac{\text{Häufigkeit von } (a_k, b_l)}{\text{Gesamtgröße der Stichprobe}}$$

- Die relativen Häufigkeiten h_{kl} aller vorkommenden Merkmalsausprägungen (a_k, b_l) bilden wieder die Gewichte einer *Wahrscheinlichkeitsverteilung*, d.h. $h_{kl} \geq 0$, und

$$\sum_k \sum_l h_{kl} = 1$$

Kontingenztabelle

Definition

In einer **Kontingenztabelle** sind die Häufigkeiten der kombinierten Merkmalsausprägungen, sowie, in den Randspalten, die Häufigkeiten der Ausprägungen der einzelnen Merkmale dargestellt:

X/Y	b_1	b_2	$\dots$	b_s	Summe
a_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,s}$	n_1^X
a_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,s}$	n_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
a_r	$n_{r,1}$	$n_{2,r}$	$\cdots$	$n_{r,s}$	n_r^X
Summe	n_1^Y	n_2^Y	$\cdots$	n_s^Y	n

- Falls sowohl X als auch Y quantitative Merkmalsausprägungen haben, spricht man auch von einer Korrelationstabelle.
- Sind die Merkmalsausprägungen von X und/oder Y kontinuierlich, oder ist die Anzahl der möglichen Merkmalsausprägungen zu groß, dann unterteilt man die Merkmalsausprägungen zunächst in Klassen, und erstellt dann eine Kontingenz- bzw. Korrelationstabelle für die klassierten Daten.
- Die absoluten Häufigkeiten n_k^X und n_l^Y der einzelnen Merkmale X und Y sind gegeben als

$$n_k^X = \sum_l n_{k,l} \qquad \text{sowie} \qquad n_l^Y = \sum_k n_{k,l}$$

wobei sich die Summe für n_k^X über die k-te Zeile (also von 1 bis s) und für n_l^Y über die l-te Spalte (von 1 bis r) erstreckt.

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)

Hair/Eye	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	*68*	*20*	*15*	*5*	*108*
Brown	*119*	*84*	*54*	*29*	*286*
Red	*26*	*17*	*14*	*14*	*71*
Blond	*7*	*94*	*10*	*16*	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

Beispiel. (Tageshöchsttemperaturen in Bonn)

Temp./Monat	*Mai*	*Juni*	*Juli*	*Aug.*	*Sept.*	$\sum$
≤ 15	*24*	*3*	*0*	*1*	*10*	*38*
(15,20]	*5*	*15*	*2*	*9*	*10*	*41*
(20,25]	*1*	*7*	*19*	*7*	*5*	*39*
>25	*0*	*5*	*10*	*14*	*5*	*34*
$\sum$	*30*	*30*	*31*	*31*	*30*	*152*

- Man kann auch Kontingenztabellen für mehr als zwei Merkmale aufstellen. Da die Tabellen dann mindestens dreidimensional sind, hält man zur Darstellung die Ausprägungen eines oder mehrerer Merkmale fest, und zeigt die entsprechenden Teiltabellen.
- In den folgenden Teiltabellen der Haar- und Augenfarben von Statistikstudenten ist das Merkmal Geschlecht festgehalten:

Female:

Hair/Eye	Brown	Blue	Hazel	Green	$\sum$
Black	36	9	5	2	52
Brown	66	34	29	14	143
Red	16	7	7	7	37
Blond	4	64	5	8	81
$\sum$	122	114	46	31	313

Male:

Hair/Eye	Brown	Blue	Hazel	Green	$\sum$
Black	32	11	10	3	56
Brown	53	50	25	15	143
Red	10	10	7	7	34
Blond	3	30	5	8	46
$\sum$	98	101	47	33	279

- Die Kontingenztabelle kann man auch für die relativen Häufigkeiten bilden. Dabei wird jeder Eintrag durch n geteilt.

$$h_{k,l} = n_{k,l}/n, \quad h_k^X = n_k^X/n, \quad h_l^Y = n_l^Y/n$$

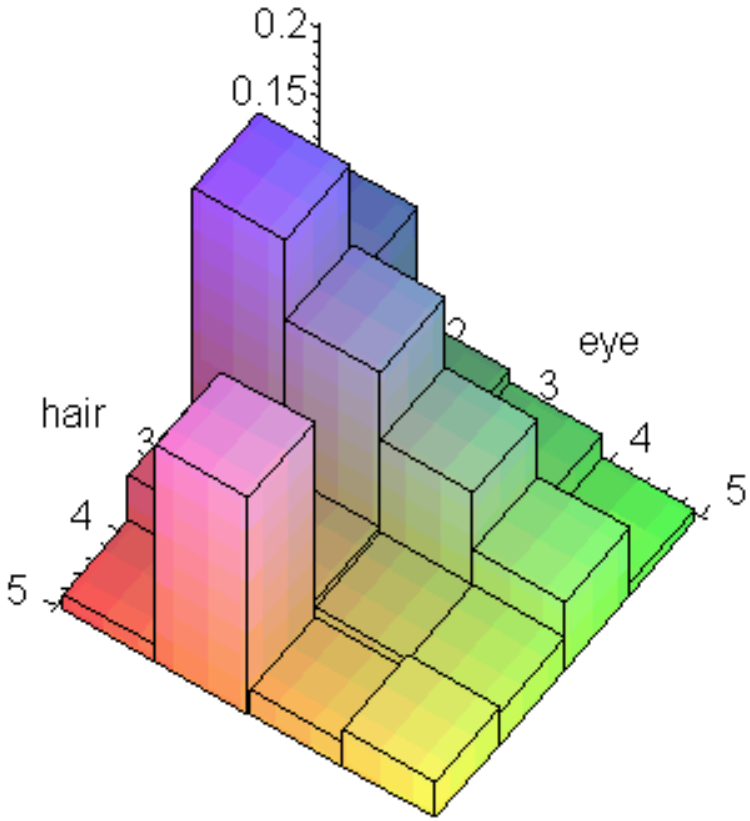
X/Y	b_1	b_2	$\dots$	b_s	Summe
a_1	$h_{1,1}$	$h_{1,2}$	$\cdots$	$h_{1,s}$	h_1^X
a_2	$h_{2,1}$	$h_{2,2}$	$\cdots$	$h_{2,s}$	h_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
a_r	$h_{r,1}$	$h_{2,r}$	$\cdots$	$h_{r,s}$	h_r^X
Summe	h_1^Y	h_2^Y	$\cdots$	h_s^Y	1

Kontingenztabelle

Zweidimensionale empirische Verteilung

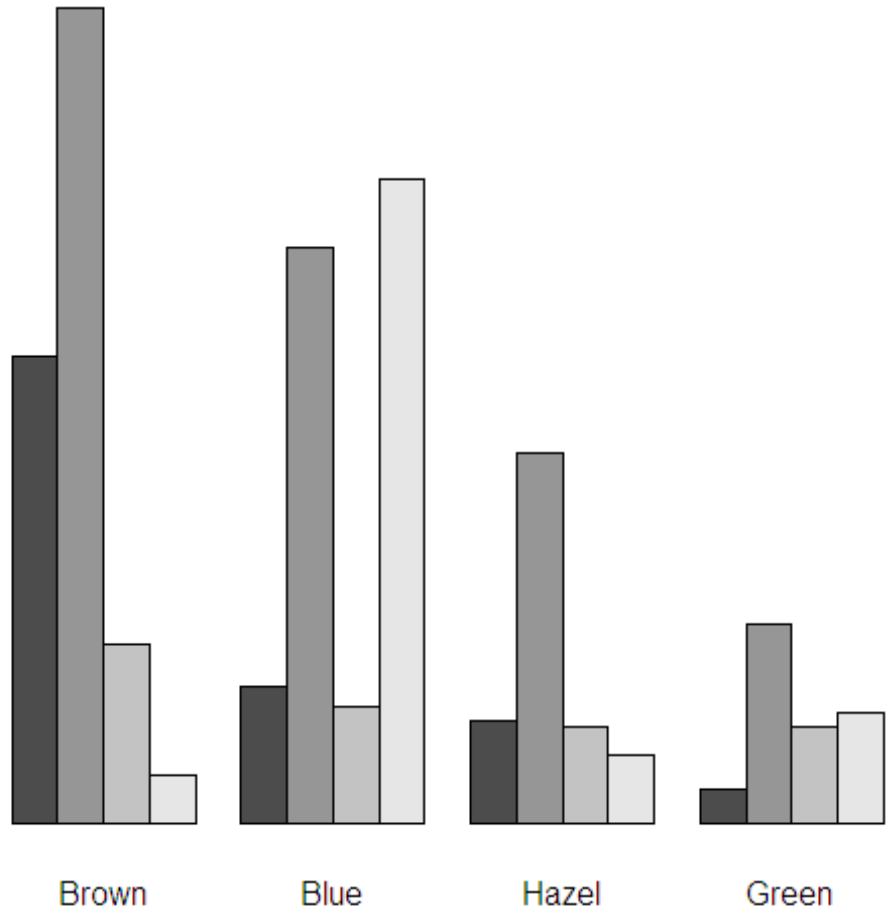
- Die in der Kontingenztabelle dargestellten relativen Häufigkeiten h_{kl} bilden die **empirische Verteilung** des zweidimensionalen Merkmals (X, Y).
- Graphisch kann man die Kontingenztabelle bzw. die zweidimensionale empirische Verteilung als **zweidimensionales Stabdiagramm** darstellen.
- Eine solche Darstellung ist allerdings etwas unübersichtlich. Alternativ kann man auch die **Stabdiagramme** der Häufigkeitsverteilungen des Merkmals X bei fest gehaltener Merkmalsausprägung von Y nebeneinaderzeichnen. Andere graphische Darstellungen für zweidimensionale empirische Verteilungen betrachten wir weiter unten.

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)



Hair/Eye	*Brown*	*Blue*	*Hazel*	*Green*
Black	*0.11*	*0.03*	*0.03*	*0.01*
Brown	*0.20*	*0.14*	*0.09*	*0.05*
Red	*0.04*	*0.03*	*0.02*	*0.02*
Blond	*0.01*	*0.16*	*0.02*	*0.03*

Stabdiagramme der Haarfarben bei festgehaltener Augenfarbe:



Kontingenztabelle

Randverteilungen

Die summierten relativen Häufigkeiten

$$h_k^X = \sum_l h_{k,l} \qquad \text{sowie} \qquad h_l^Y = \sum_k h_{k,l}$$

(wobei sich die Summe für h_k^X über die k-te Zeile und für h_l^Y über die l-te Spalte der Kontingenztabelle erstreckt)
sind genau die Gewichte der empirischen Verteilungen der einzelnen Merkmale X bzw. Y.

Definition

Die empirischenVerteilungen der einzelnen Merkmale X und Y heißen **Randverteilungen** der zweidimensionalen empirischen Verteilung h_{kl}.

Die Randverteilungen heißen so, weil sie am Rand der Kontingenztabelle stehen:

X / Y	b_1	b_2	$\dots$	b_s	Summe
a_1	$h_{1,1}$	$h_{1,2}$	$\cdots$	$h_{1,s}$	h_1^X
a_2	$h_{2,1}$	$h_{2,2}$	$\cdots$	$h_{2,s}$	h_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
a_r	$h_{r,1}$	$h_{2,r}$	$\cdots$	$h_{r,s}$	h_r^X
Summe	h_1^Y	h_2^Y	$\cdots$	h_s^Y	1

Rel. Häufigkeiten der emp. X-Verteilung

Rel. H'keiten der emp. Y-Verteilung

Bei einer Kontingenztabelle für absolute Häufigkeiten stehen entsprechend die absoluten Häufigkeiten der einzelnen Merkmale X und Y am Rand:

X / Y	b_1	b_2	$\dots$	b_s	Summe
a_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,s}$	n_1^X
a_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,s}$	n_2^X
$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
a_r	$n_{r,1}$	$n_{2,r}$	$\cdots$	$n_{r,s}$	n_r^X
Summe	n_1^Y	n_2^Y	$\cdots$	n_s^Y	n

Häufigkeiten der emp. X-Verteilung

H'keiten der emp. Y-Verteilung

Eine interaktive Demonstration zu Randverteilungen findet man im Internet unter
http://demonstrations.wolfram.com/DiscreteMarginalDistributions

3.2. Bedingte empirische Verteilungen

- Häufig interessiert uns, ob und wie ein Merkmal X mit einem anderen Merkmal Y zusammenhängt.
- Um dies zu erkennen, betrachten wir die empirische Verteilung von X unter der Bedingung, daß Y eine bestimmte Ausprägung b_l hat (und entsprechend umgekehrt).
- Dazu beschränken wir uns auf diejenigen statistischen Einheiten ω_i, für die $y_i = b_l$ gilt, d.h. wir betrachten nur die Häufigkeiten in der **l-ten Spalte der Kontingenztabelle** !

Definition

Die **bedingte relative Häufigkeit** einer Ausprägung a_k des Merkmals X gegeben die Ausprägung b_l des Merkmals Y ist

$$h_{k|l} = h(X = a_k | Y = b_l) = \frac{n_{kl}}{n_l^Y} = \frac{h_{kl}}{h_l^Y}$$

Definition

Die **bedingte relative Häufigkeit** einer Ausprägung a_k des Merkmals X gegeben die Ausprägung b_l des Merkmals Y ist

$$h_{k|l} = h(X = a_k | Y = b_l) = \frac{n_{kl}}{n_l^Y} = \frac{h_{kl}}{h_l^Y}$$

- Wenn wir die Ausprägung b_l des Merkmals Y festhalten, dann bilden die bedingten relativen Häufigkeiten $h_{k|l}$, $k = 1, 2, \ldots, r$, der möglichen Ausprägungen des Merkmals X wieder die Gewichte einer Wahrscheinlichkeitsverteilung, d.h., sie summieren sich zu 1. Diese Wahrscheinlichkeitsverteilung heißt **bedingte empirische Verteilung** des Merkmals X gegeben $Y = b_l$.
- Die bedingte empirische Verteilung von X gegeben $Y = b_l$ ist also die Verteilung, die sich aus den Häufigkeiten in der **l-ten Spalte** der Kontingenztabelle berechnet.

- Entsprechend berechnet man die bedingte empirische Verteilung von Y gegeben $X = a_k$ aus den Häufigkeiten in der k-**ten Zeile** der Kontingenztabelle:

$$\widetilde{h}_{l|k} = h(Y = b_l | X = a_k) = \frac{n_{kl}}{n_k^X} = \frac{h_{kl}}{h_k^X}$$

- **Beachte:**

$$h(X = a_k | Y = b_l) \neq h(Y = b_l | X = a_k) \quad \textbf{!!!}$$

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)

Hair / Eye	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	*68*	*20*	*15*	*5*	*108*
Brown	*119*	*84*	*54*	*29*	*286*
Red	*26*	*17*	*14*	*14*	*71*
Blond	*7*	*94*	*10*	*16*	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

Bedingte emp. Verteilungen von Haarfarbe gegeben Augenfarbe:

Hair | Eye	*Brown*	*Blue*	*Hazel*	*Green*
Black	$\frac{68}{220}$	$\frac{20}{215}$	$\frac{15}{93}$	$\frac{5}{64}$
Brown	$\frac{119}{220}$	$\frac{84}{215}$	$\frac{54}{93}$	$\frac{29}{64}$
Red	$\frac{26}{220}$	$\frac{17}{215}$	$\frac{14}{93}$	$\frac{14}{64}$
Blond	$\frac{7}{220}$	$\frac{94}{215}$	$\frac{10}{93}$	$\frac{16}{64}$
$\sum$	1	1	1	1

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)

Eye | *Hair*	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	*68*	*20*	*15*	*5*	*108*
Brown	*119*	*84*	*54*	*29*	*286*
Red	*26*	*17*	*14*	*14*	*71*
Blond	*7*	*94*	*10*	*16*	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

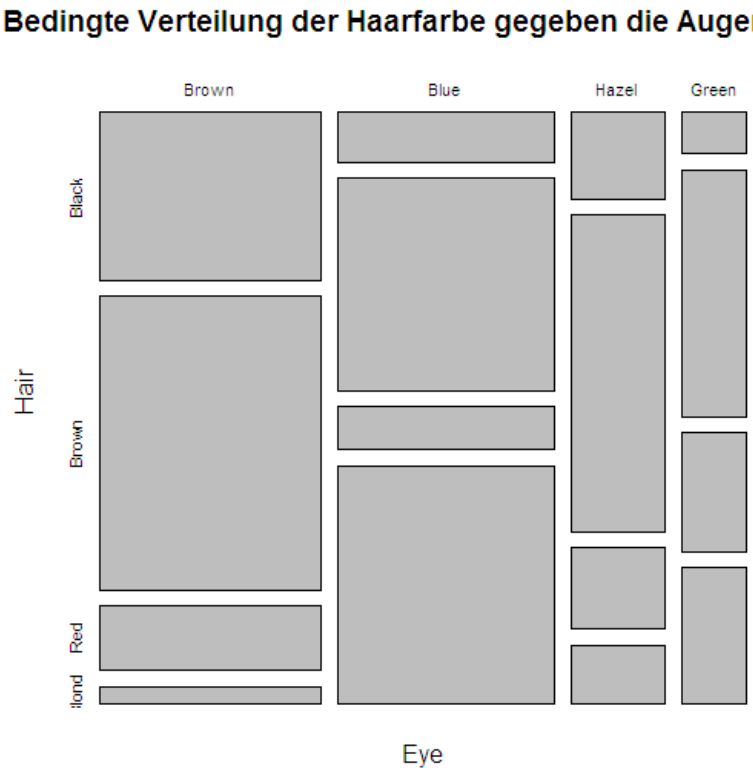
Bedingte emp. Verteilungen von Augenfarbe gegeben Haarfarbe:

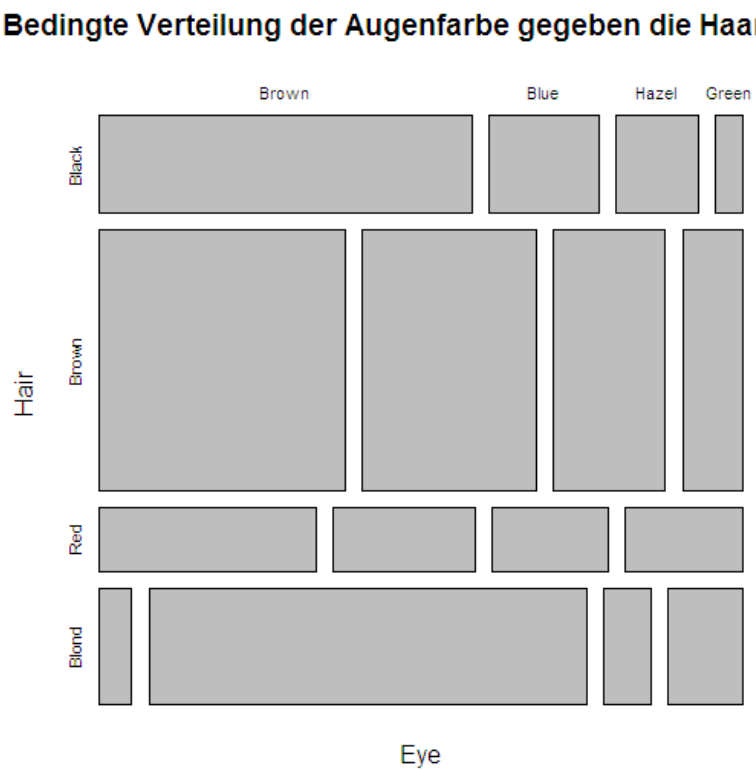
Eye/Hair	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	$\frac{68}{108}$	$\frac{20}{108}$	$\frac{15}{108}$	$\frac{5}{108}$	1
Brown	$\frac{119}{286}$	$\frac{84}{286}$	$\frac{54}{286}$	$\frac{29}{286}$	1
Red	$\frac{26}{71}$	$\frac{17}{71}$	$\frac{14}{71}$	$\frac{14}{71}$	1
Blond	$\frac{7}{127}$	$\frac{94}{127}$	$\frac{10}{127}$	$\frac{16}{127}$	1

Bedingte empirische Verteilungen

Mosaikplot

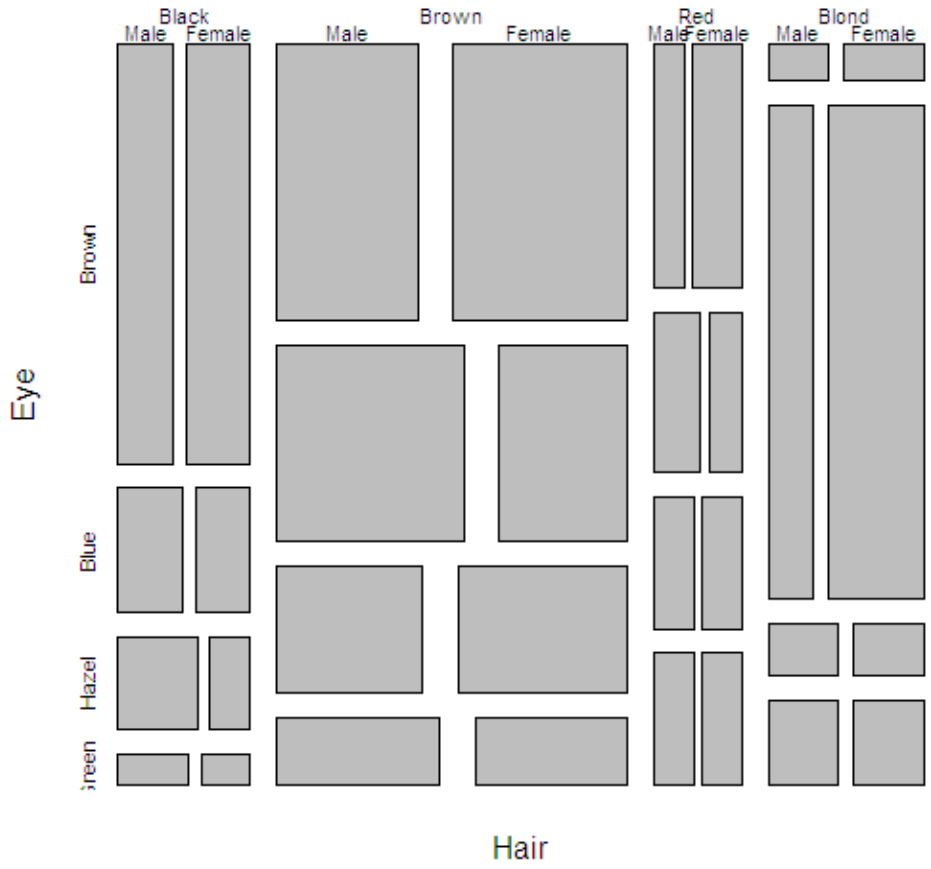
Graphisch stellt man die bedingten empirischen Verteilungen eines Merkmals gegeben die verschiedenen Merkmalsausprägungen eines anderen Merkmals nebeneinander in einem **Mosaikplot** dar:





Man erkennt, daß Haar- und Augenfarbe nicht unabhängig sind.

Dagegen zeigt eine weitere Aufspaltung nach dem Geschlecht, daß die Kombination Haar-/Augenfarbe nicht sehr stark von diesem abhängt. Allerdings ist die Merkmalsausprägung "braunhaarig mit blauen Augen" anscheinend unter Männern überrepräsentiert, wohingegen die Merkmalsausprägung "blond mit blauen Augen" eher bei Frauen auftritt:



3.3. Maße für die statistische Abhängigkeit

Unabhängige Merkmale

- Wenn zwei Merkmale X und Y nicht zusammenhängen, dann sollte die empirische Verteilung von X in einer "genügend großen" ("repräsentativen") Stichprobe kaum noch von der Merkmalsausprägung von Y abhängen.
- Für beliebige Ausprägungen a_k, b_l, und $\widetilde{b_l}$ sollte also gelten:

$$\begin{aligned} h_{k|l} &= h(X = a_k | Y = b_l) \approx h(X = a_k | Y = \widetilde{b_l}) \\ &\approx h(X = a_k) = h_k^X \qquad (*) \end{aligned}$$

- Wegen $h_{k|l} = h_{kl} / h_l^Y$ ist die Bedingung (*) aber gerade dann erfüllt, wenn

$$h_{kl} \approx h_k^X \cdot h_l^Y \qquad \text{bzw.} \qquad \frac{h_{kl}}{h_k^X \cdot h_l^Y} \approx 1 \qquad (**)$$

für $k = 1, 2, \ldots, r$ und $l = 1, 2, \ldots, s$ gilt.

Maße für die statistische Abhängigkeit

Quadratische Kontingenz

Beobachtung: *Für X unabhängig von Y sollte* $\frac{h_{kl}}{h_k^X \cdot h_l^Y} \approx 1$ *gelten.*

- Also liegt es nahe, die mittlere quadratische Abweichung des Quotienten von 1 als Maß für die Abhängigkeit der Daten zu betrachten:

Definition

1. Die **mittlere quadratische Kontingenz** der Stichprobe ist

$$\phi^2 = \sum_{k,l} \left(\frac{h_{kl}}{h_k^X h_l^Y} - 1 \right)^2 h_k^X h_l^Y = \sum_{k,l} \frac{\left(h_{kl} - h_k^X h_l^Y \right)^2}{h_k^X h_l^Y}$$

2. Die χ^2**-Statistik** oder **quadratische Kontingenz** ist

$$\chi^2 = \sum_{k,l} \frac{(n_{kl} - \widehat{n}_{kl})^2}{\widehat{n}_{kl}} \qquad \text{mit} \qquad \widehat{n}_{kl} = \frac{1}{n} n_k^X n_l^Y \ .$$

- Nachrechnen ergibt sofort:

$$\phi^2 = \frac{\chi^2}{n} = \frac{1}{n} \sum_{k,l} \frac{(n_{kl} - \widehat{n}_{kl})^2}{\widehat{n}_{kl}}$$

- Hierbei ist

$$\widehat{n}_{kl} = \frac{1}{n} n_k^X \, n_l^Y$$

die absolute Häufigkeit der kombinierten Merkmalsausprägung (a_k, b_l), welche man in etwa erwarten würde, wenn die Merkmale X und Y voneinander unabhängig wären.

- Mithilfe der χ^2-Statistik können wir die mittlere quadratische Kontingenz also auch direkt aus den *absoluten* Häufigkeiten, d.h. aus der Kontingenztabelle berechnen. Dazu gehen wir wie folgt vor:

1. Wir stellen zunächst eine Tabelle für die bei Unabhängigkeit zu erwartenden Häufigkeiten $\widehat{n}_{kl}$ auf.
2. Anschließend berechnen wir aus dieser Tabelle und der Kontingenztabelle die Statistiken χ^2 und ϕ^2.

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)

n_{kl}	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	*68*	*20*	*15*	*5*	*108*
Brown	*119*	*84*	*54*	*29*	*286*
Red	*26*	*17*	*14*	*14*	*71*
Blond	*7*	*94*	*10*	*16*	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

$$\chi^2 = 138.3,\ \phi^2 = 0.23$$

$\widehat{n}_{kl}$	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	$\frac{220*108}{592} = 40.1$	$\frac{215*108}{592} = 39.2$	17.0	11.7	*108*
Brown	$\frac{220*286}{592} = 106.3$	$\frac{215*286}{592} = 103.9$	44.9	30.9	*286*
Red	$\frac{220*71}{592} = 26.4$	$\frac{215*71}{592} = 25.8$	11.2	7.7	*71*
Blond	$\frac{220*127}{592} = 47.2$	$\frac{215*127}{592} = 46.1$	20.0	13.7	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

- *Wie sollen wir das Ergebnis $\phi^2 = 0.23$ interpretieren ? Bedeutet es eher eine starke oder eine schwache Abhängigkeit der Haarfarbe von der Augenfarbe ?*
- Um diese Frage zu beantworten, müssen wir uns überlegen, in welchem Bereich die Werte der mittleren quadratischen Kontingenz ϕ^2 liegen können.
- Sind die Merkmale X und Y **unabhängig**, so gilt näherungsweise $h_{kl} = h_k^X \cdot h_l^Y$, also $\phi^2 \approx 0$.
- Läßt sich hingegen die Ausprägung von **Y aus** der Ausprägung von **X vollständig vorhersagen und** ebenso **umgekehrt**, dann hat die Kontingenztabelle in jeder Zeile und Spalte nur einen von 0 verschiedenen Eintrag, und es gilt

$$\phi^2 = \min(r - 1, s - 1),$$

wobei r und s die Spalten- und Zeilenzahl der Kontingenztabelle ist.

Maße für die statistische Abhängigkeit

Quadratische Kontingenz

- *Fazit:* Die Werte von ϕ^2 liegen zwischen 0
  (was auf Unabhängigkeit der Merkmale X und Y hindeutet),
  und $\min(r-1, s-1)$
  (was auf vollständige Abhängigkeit von X und Y hinweist).
  Daher definiert man:

Definition

Die **normierte Kontingenz** (das **Cramérsche Kontingenzmaß**) ist

$$C = \sqrt{\frac{\phi^2}{\min(r-1, s-1)}} = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, s-1)}}$$

- Die normierte Kontingenz nimmt Werte zwischen 0 und 1 an.

Beispiel. (Haar- und Augenfarbe von Statistikstudenten)

- *In unserem Beispiel erhalten wir:*

n_{kl}	*Brown*	*Blue*	*Hazel*	*Green*	$\sum$
Black	*68*	*20*	*15*	*5*	*108*
Brown	*119*	*84*	*54*	*29*	*286*
Red	*26*	*17*	*14*	*14*	*71*
Blond	*7*	*94*	*10*	*16*	*127*
$\sum$	*220*	*215*	*93*	*64*	*592*

$$\chi^2 = 138.3,\ \phi^2 = 0.23,$$

$$\Longrightarrow \qquad C = \sqrt{\frac{0.23}{3}} = 0.28$$

- *Es existiert also ein Zusammenhang zwischen der Augen- und Haarfarbe - dieser ist aber nicht sehr stark ausgeprägt, da* $C < 0.5$ *gilt.*

3.4. Kovarianz und Korrelation

- Vor allem bei kontinuierlichen Merkmalen, aber auch bei diskreten Merkmalen mit großem Wertebereich ist die Kontingenztabelle i.d.R. hauptsächlich mit Nullen und Einsen gefüllt.
- Beispielsweise betrachten wir eine Urliste mit Werbeausgaben und Verkaufszahlen einer Firma:

Monat	1	2	3	4	5	6	7	8	9	10
Werbeausgaben	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1
Verkäufe	101	92	110	120	90	82	93	75	91	105

- Wir wollen die Daten in einer Kontingenztabelle abbilden, bei der sowohl die X als auch die Y-Unterteilung sehr fein ist:

Kovarianz und Korrelation

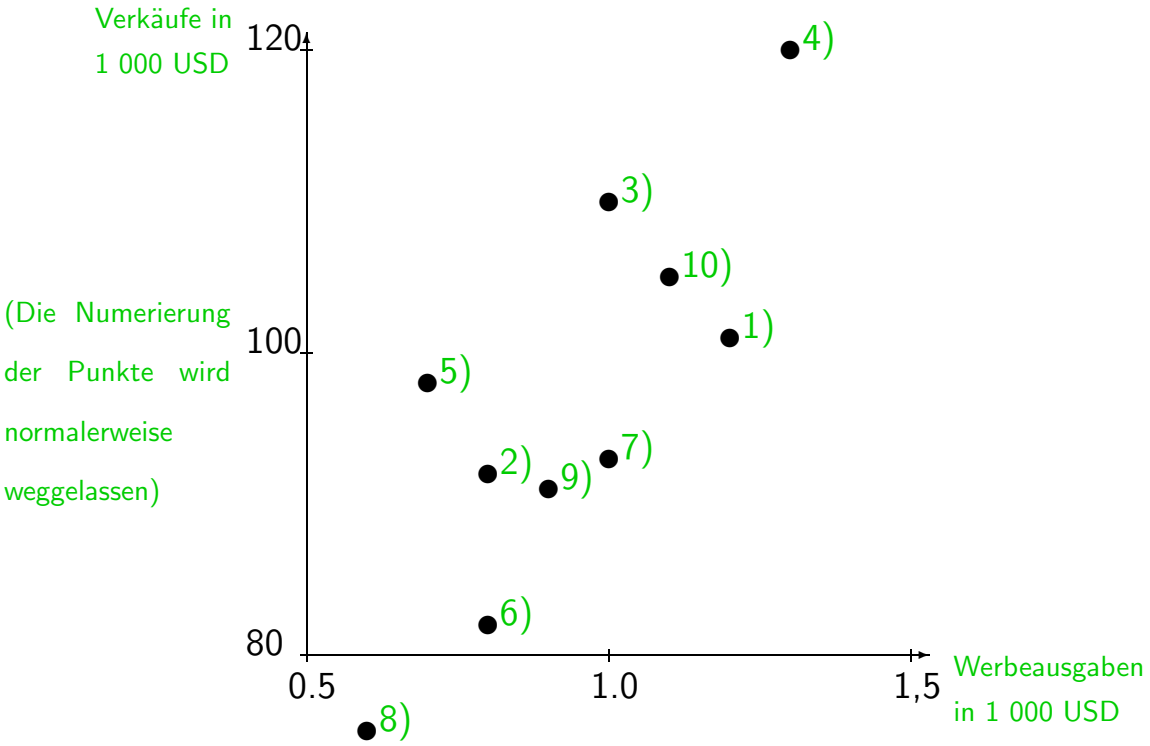
Streudiagramm

82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(Wert (1.3,120) und (1.0,110) und (0.6,75) nicht dargestellt)

- In diesem Fall ist die Kontingenztabelle sehr unübersichtlich!
- Natürlich könnten wir die Daten in eine kleine Anzahl von Klassen einteilen, aber dadurch würde viel Information verlorengehen.
- Stattdessen empfiehlt sich hier die Darstellung im **Streudiagramm**:

- In diesem Fall empfiehlt sich eher die Darstellung im Streudiagramm:



Monat	1	2	3	4	5	6	7	8	9	10
Werbeausgaben	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1
Verkäufe	101	92	110	120	90	82	93	75	91	105

Kovarianz und Korrelation

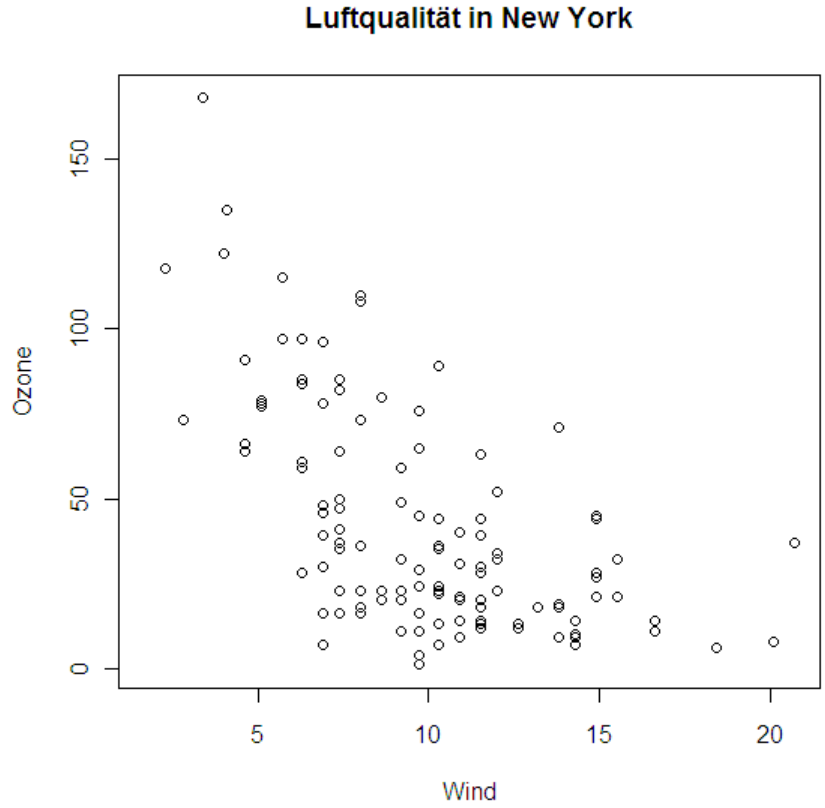
Streudiagramm, positive Korrelation

Definition

In einem **Streudiagramm** wird eine Urliste mit zwei quantitativen Merkmalen dargestellt, indem die Wertepaare in ein zweidimensionales Koordiantensystem eingezeichnet werden.

- Oft kann man sich aus der qualitativen Betrachtung der Punktwolke bereits ein gutes Bild über den Zusammenhang der beiden Merkmale machen.
- Beispielsweise sieht man bei Betrachtung der Form der Punktwolke im Streudiagramm von oben ein Ansteigen der Verkaufswerte Y mit wachsenden Werbeausgaben X, d.h. die Merkmale X und Y sind **positiv korreliert**.
- Dieses "Ansteigen" ist aber nur die allgemeine Tendenz. Es gibt natürliche Fluktuationen, d.h einzelne Werte (z.B. 5) und 9)), bei denen die höheren Werbeausgaben nicht mit einem höheren Verkaufsergebnis zusammenfallen.

Beispiel. (Negative Korrelation)



In diesem Streudiagramm sind Meßwerte der Ozonkonzentration und der Windstärke an Sommertagen in New York aufgetragen. Man erkennt deutlich, daß die Ozonkonzentration bei größerer Windstärke eher kleiner ist - die beiden Merkmale sind **negativ korreliert**.

- Umgangssprachlich nennt man zwei quantitative Merkmale X und Y positiv korreliert, falls eine Zunahme beim Merkmal X statistisch gesehen zu einer Zunahme beim Merkmal Y führt.
- Entsprechend heißen X und Y negativ korreliert, falls eine Zunahme beim Merkmal X statistisch gesehen zu einer Abnahme beim Merkmal Y führt.
- Wir wollen nun die Größe der Korrelation zwischen zwei Merkmalen quantifizieren.

Beispiel. (Macht es Sinn, Werbung zu betreiben ?)

Angenommen sei, daß Werbung hauptsächlich kurzfristig wirkt, d.h. nur in dem Monat, in dem die Werbekampagne läuft (diese Annahme ist nicht unumstritten). Wir betrachten die monatlichen Werbeausgaben und die Verkaufszahlen (in 1000 USD) in großen Läden für Farmerbedarf in den USA für eine Stichprobe von 10 Monaten:

Monat	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Werbeausgaben	*1.2*	*0.8*	*1.0*	*1.3*	*0.7*	*0.8*	*1.0*	*0.6*	*0.9*	*1.1*
Verkäufe	*101*	*92*	*110*	*120*	*90*	*82*	*93*	*75*	*91*	*105*

- *Wir fragen: Wird in den Monaten, in denen mehr geworben wird, auch mehr verkauft? Mit anderen Worten, sind die Verkäufe und die Werbeausgaben positiv korreliert? Und wenn ja, wie stark ist der Zusammenhang?*

Kovarianz und Korrelation

Empirische Kovarianz

- Ein Maß für den Zusammenhang zweier quantitativer Merkmale X und Y ist die empirische Kovarianz:

Definition

Seien $(x_1, y_1), \ldots, (x_n, y_n)$ die beobachteten kombinierten Merkmalsausprägungen der quantitativen Merkmale X und Y (also etwa als Klassenmittelpunkte). Dann ist die empirische Kovarianz der Beobachtungswerte gegeben durch

$$c_{x,y} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- Gilt $c_{x,y} > 0$, dann sind die Beobachtungswerte positiv korreliert.
- Gilt $c_{x,y} = 0$, dann sind die Beobachtungswerte unkorreliert.
- Gilt $c_{x,y} < 0$, dann sind die Beobachtungswerte negativ korreliert.

$$c_{x,y} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- *Abweichende Konvention:* Betrachtet man Stichproben von Grundgesamtheiten, dann teilt man (wie schon bei der Definition der empirischen Varianz) häufig durch $n - 1$ statt durch n.
- *Zusammenhang mit der empirischen Varianz:* Die empirische Varianz der Beobachtungswerte $x_1, \ldots, x_n$ ist

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = c_{x,x}$$

Kovarianz und Korrelation

Berechnung der empirischen Kovarianz aus relativen Häufigkeiten

Sind die Daten klassiert, und sind n_{kl} und h_{kl} die relativen Häufigkeiten der Klassen (siehe Kontingenztabelle), dann gilt näherungsweise

$$c_{x,y} \approx \sum_k \sum_l h_{kl} \cdot (a_k - \overline{x}) \cdot (b_l - \overline{y}) = \frac{1}{n} \sum_k \sum_l n_{kl} \cdot (a_k - \overline{x}) \cdot (b_l - \overline{y})$$

Hierbei sind $a_1, \ldots, a_r$ die Klassenmittelpunkte der Klasseneinteilung für die Ausprägungen von X, $b_1, \ldots, b_s$ sind die Klassenmittelpunkte für Y, und

$$\overline{x} \approx \frac{1}{n} \sum_k n_k^X \cdot a_k, \qquad \overline{y} \approx \frac{1}{n} \sum_l n_l^Y \cdot b_l$$

sind die arithmetischen Mittelwerte, die wir ebenfalls näherungsweise aus den Häufigkeiten der Klassen berechnen können.

- Ähnlich wie für die Varianz können wir folgende praktische Formel für die Kovarianz ableiten:

$$\boxed{c_{x,y} = \overline{xy} - \bar{x}\bar{y}}$$

Dabei benutzt man folgende Formeln für die Mittelwerte:

- Für Daten aus der Urliste:

$$\bar{x} = (1/n)\sum_i x_i,\ \bar{y} = (1/n)\sum_i y_i, \quad \text{und} \quad \overline{xy} = (1/n)\sum_i x_i y_i\ .$$

- Für die (klassierten) Daten aus der Kontingenztabelle:

$$\begin{aligned} \bar{x} &\approx (1/n)\sum_k n_k^X a_k = \sum h_k^X a_k, \quad \bar{y} \approx (1/n)\sum_l n_l^Y b_l = \sum_l h_l^Y b_l \\ \overline{xy} &\approx (1/n)\sum_k\sum_l n_{k,l} a_k b_l = \sum_k\sum_l h_{k,l} a_k b_l\ . \end{aligned}$$

Machen wir in unserem Beispiel die Probe, ob die Daten die Wirksamkeit von Werbung hergeben!

Monat	1	2	3	4	5	6	7	8	9	10	Summe
x=Ausg.	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1	9.4
y=Verk.	101	92	110	120	90	82	93	75	91	105	959
xy	121.2	73.6	110.0	156.0	56.0	65.6	93.0	45.0	81.9	115.5	924.8

- Also erhalten wir für die Kovarianz:

$$c_{x,y} = (\overline{xy} - \bar{x}\bar{y}) = \frac{924.8}{10} - \frac{9.4}{10} \times \frac{959}{10} = 2.33$$

- Werbung und Verkäufe sind also positiv korreliert.
- Die Frage bleibt, ob das eine eher starke Korrelation ist oder eine eher schwache.

Kovarianz und Korrelation

Korrelationskoeffizient

Definition

Der Korrelationskoeffizient $\rho_{x,y}$ der beobachteten Daten ist definiert als

$$\rho_{x,y} = \frac{c_{x,y}}{\sigma_x \sigma_y}$$

- Hierbei sind σ_x und σ_y die Standardabweichungen der Beobachtungswerte $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$.
- Es gilt $-1 \leq \rho_{X,Y} \leq 1$.

- Der Korrelationskoeffizient nimmt also Werte an zwischen -1 (*perfekte negative Korrelation*) und $+1$ (*perfekte positive Korrelation*).
- Bei $\rho_{X,Y} = 0$ sind die beobachteten Merkmalsausprägungen unkorreliert.
- Eine grobe Einteilung ist die folgende:
  - $0 < |\rho_{X,Y}| < 0.5$: Schwache (neg bzw. pos.) Korrelation
  - $0.5 \leq |\rho_{X,Y}| < 0.8$: Mittlere (neg. bzw. pos.) Korrelation
  - $0.8 \leq |\rho_{X,Y}| \leq 1$: Starke (neg. bzw. pos.) Korrelation

Wie stark ist die Korrelation in unserem Beispiel? Wir erweitern die Tabelle

Monat	1	2	3	4	5	6	7	8	9	10	Summe
x=W.Ausgaben	1.2	0.8	1.0	1.3	0.7	0.8	1.0	0.6	0.9	1.1	9.4
y=Verkäufe	101	92	110	120	90	82	93	75	91	105	959
xy	121.2	73.6	110.0	156.0	56.0	65.6	93.0	45.0	81.9	115.5	924.8
x^2	1.44	0.64	1.00	1.69	0.49	0.64	1.00	0.36	0.81	1.21	9.28
y^2	10 201	8 464	12 100	14 400	8 100	6 724	8 649	5 625	8 281	11 025	93 569

- Also

$$s_X = \sqrt{\left(\frac{9.28}{10} - \left(\frac{9.4}{10}\right)^2\right)} = 0.21 \text{ und } s_Y = \sqrt{\left(\frac{93569}{10} - \left(\frac{959}{10}\right)^2\right)} = 12.7$$

- Und so ergibt sich der Korrelationskoeffizient

$$\rho_{X,Y} = \frac{2.33}{0.21 \times 12.7} = 0.88$$

- Die Korrelation ist stark, die Schwankungen im Verkauf können zu großen Teilen auf die eingesetzten Werbemittel zurückgeführt werden.

Kovarianz und Korrelation

Lineare Transformationen

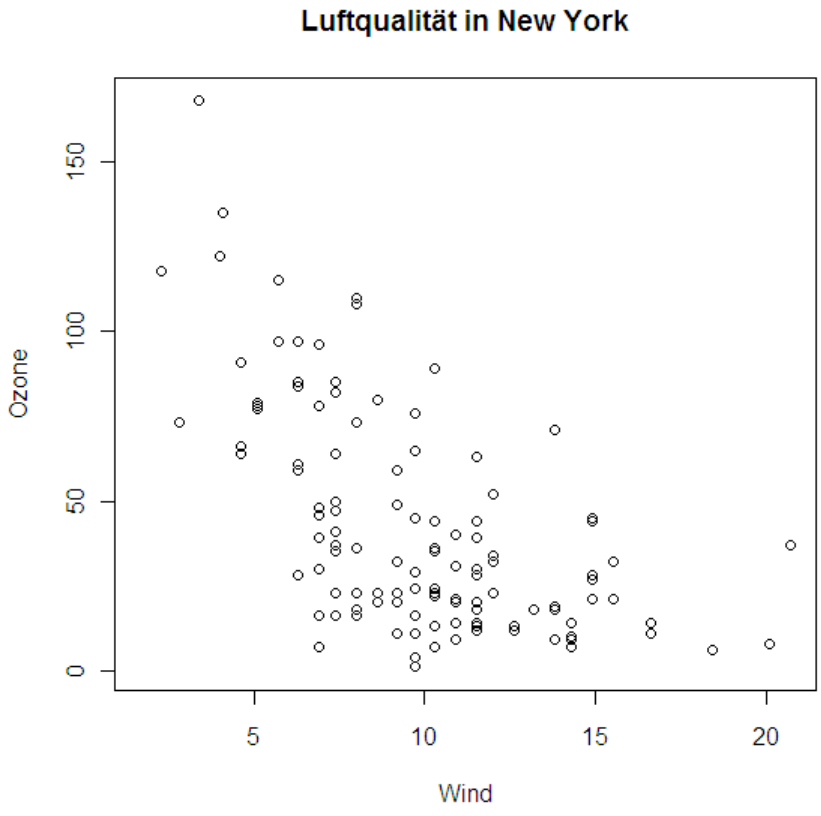
- Unter linearen Datentransformationen $x_i \longrightarrow ax_i + b$ und $y_i \longrightarrow cy_i + d$, mit $a, b \neq 0$ transformieren sich die empirischen Varianzen und Kovarianzen sowie der Korrelationskoeffizient folgendermaßen:

$$\begin{aligned} \sigma_x &\longrightarrow |a|\sigma_x \\ \sigma_y &\longrightarrow |c|\sigma_y \\ c_{x,y} &\longrightarrow a \times c \times c_{x,y} \\ \rho_{x,y} &\longrightarrow (\mathrm{sign}(a) \times \mathrm{sign}(c))\rho_{x,y} \end{aligned}$$

  Hierbei ist $sign(t) = +1$ für $t > 0$ und $sign(t) = -1$ für $t < 0$.

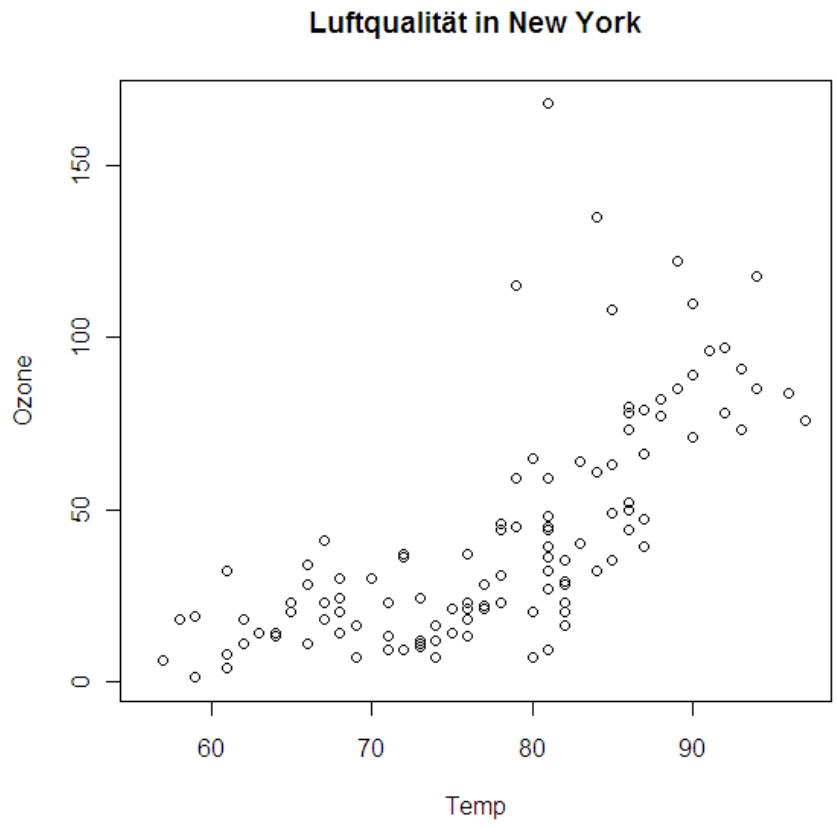
- Insbesondere hängen alle diese Größen nicht von der Nullpunktsverschiebung b und d ab.
- Der Korrelationskoeffizient ist sogar weitgehend (bis auf Vorzeichen) unverändert unter linearen Datentransformationen.

Beispiel.



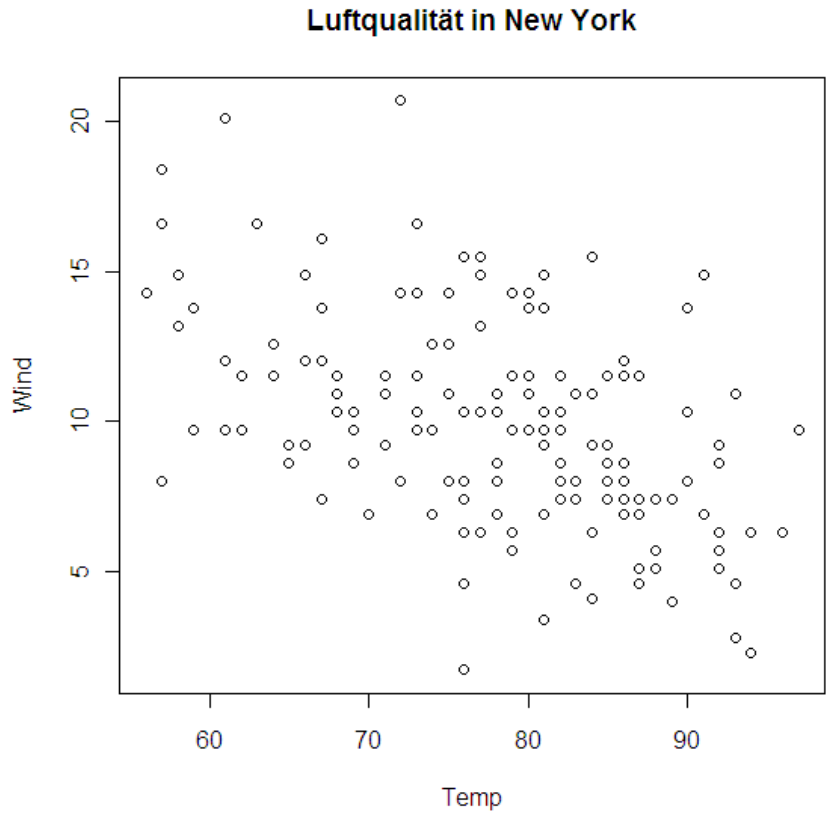
- $\rho_{x,y} = -0.60 \quad \longrightarrow \quad$ *Mittlere negative Korrelation*

Beispiel.



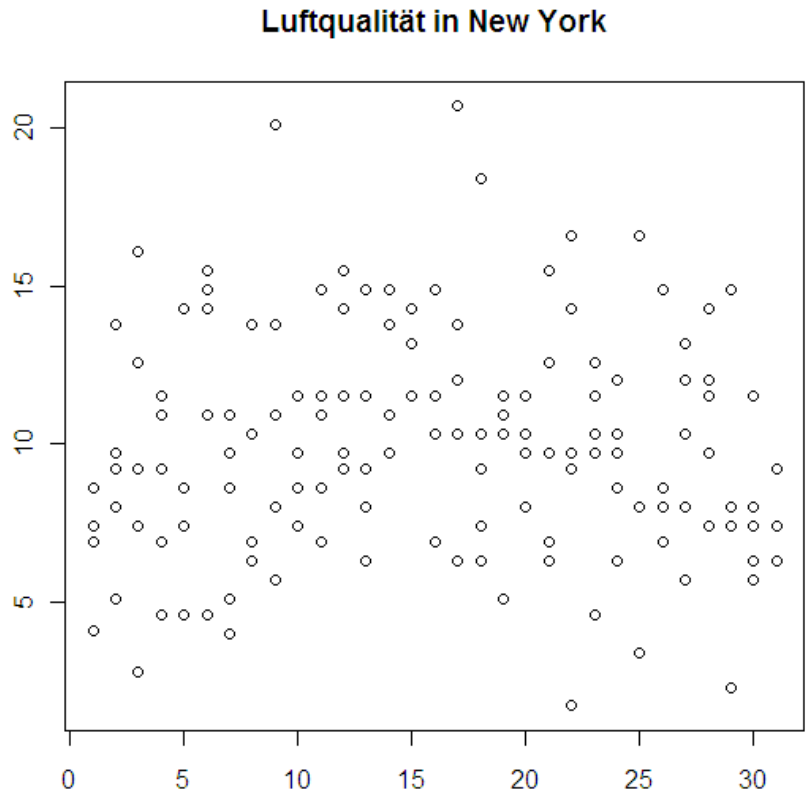
- $\rho_{x,y} = 0.70 \quad \longrightarrow \quad$ *Mittlere positive Korrelation*

Beispiel.

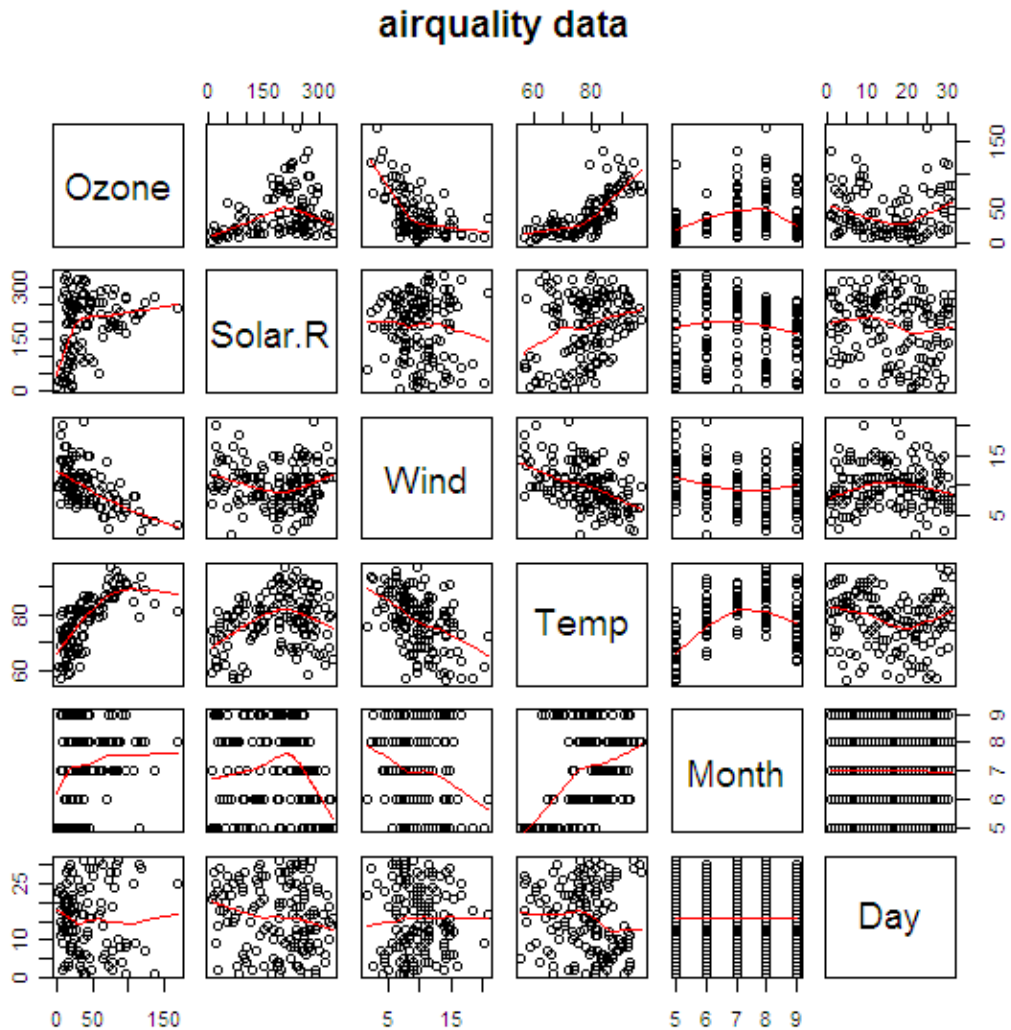


- $\rho_{x,y} = -0.46 \quad \longrightarrow \quad$ *Schwache negative Korrelation*

Beispiel.



- $\rho_{x,y} = -0.13 \quad \longrightarrow \quad$ *Sehr schwache negative Korrelation, nahezu unkorreliert*


airquality data
Ozone
Solar.R
Wind
Temp
Month
Day

Kovarianz und Korrelation

Positive Korrelation vs. Kausalzusammenhang

- Wenn der Korrelationskoeffizient $\rho_{x,y}$ nicht verschwindet, kann man dann von einer kausalen Abhängigkeit sprechen?
- In unserem Beispiel oben hatten wir z.B. gesagt "Mehr Werbung in einem Monat (X) führt zu mehr Verkauf Y in diesem Monat". Das scheint zu stimmen. Wenn wir aber die Rollen von X und Y vertauschen, bekämen wir: "Ein höherer Verkauf Y in einem bestimmten Monat führt zu höheren Werbeausgaben X in demselben Monat", was ziemlich unsinnig erscheint.
- Eine nicht verschwindende Korrelation kann also auf eine kausale Abhängigkeit hindeuten, muß aber nicht.

Kovarianz und Korrelation

Kausalmodelle bei positiver Korrelation

- Folgende Kausalmodelle sind bei positiver Korrelation möglich:

Positive Korrelation - kausale Deutungen

- Mehr X fürt zu mehr Y
- Mehr Y führt zu mehr X
- Es gibt eine verborgene Ursache Z, so dass eine Veränderung von Z zu einer gleichzeitigen Zunahme von X und Y führt.

- Welches der Kausalmodelle gültig ist, kann man nicht so leicht mit Statistik herausfinden.

Kovarianz und Korrelation

Scheinkorrelationen

- Ein Beispiel für eine versteckte Korrelation ist das folgende: Eine boomende Konjunktur (versteckte Ursache Z gemessen als Wirtschaftswachstum in %) führt zu dem Kauf von mehr im Inland produzierten HiFi-Anlagen (Merkmal X) und zum Kauf von mehr importierten HiFi-Anlagen (Merkmal Y). Wir finden also einen positiven Korrelationskoeffizienten $\rho_{X,Y}$.
- Das Kausalmodell 1) "Verstärkter Kauf von im Inland produzierten Anlagen führt zum Kauf von mehr importierten Anlagen" oder das Kausalmodell 2) "Mehr Import von HiFi Anlagen führt zum Verkauf von mehr im Inland produierten Anlagen" sind beide gleichermaßen unsinnig.
- Wenn in diesem Bsp. der direkte Zusammenhang von X und Y interessiert, muß der Z–Effekt herausgerechnet werden. Dies geschieht wieder mit einer Bedingung: Die Korrelation von X und Y gegeben ein bestimmtes Wirtschaftswachstum Z dürfte negativ ausfallen.
- Das Kausalmodell 3) heißt auch Scheinkorrelation von X und Y.

Weitere Beispiele für Scheinkorrelationen:

-Wer auf engem Raum wohnt, ist eher gewalttätig.

-Wer viel im Internet surft, ist weniger gewalttätig.

-Wer gut in Englisch ist, der ist auch gut in Physik.

-Wenn man auf die beste Schule am Ort geht, dann verdient man später viel Geld.

3.5. Lineare Regression

Wenn wir durch theoretische Überlegungen festgelegt haben, welche Größe als Ursache X (z.B. Werbung) für die Veränderung der anderen Y (Verkaufszahlen) in Frage kommt, können wir Modelle für diese Beeinflussung aufstellen.

- Für den Fall einer perfekten Abhängigkeit, erwarten wir einen funktionalen Zusammenhang der Form

$$\boxed{Y = f(X)}$$

- Hier ist f eine Funktion, die natürlich erstmal gefunden werden muß.
- Sehr oft ist jedoch die Veränderung von X nicht als die alleinige Ursache der Veränderung von Y anzusehen. Wir erhalten ein Modell:

$$\boxed{Y = f(X) + \text{ statistische Schwankungen}}$$

- Die statistischen Schwankungen stehen hier für alle möglichen von X unabhängigen Ursachen für die Veränderung von Y.

Lineare Regression

- Aufgabe der Modellbildung ist das Auffinden einer guten Funktion f.
- Eine Funktion bzw. ein Modell ist immer dann gut, wenn sie/es einfach zu handhaben ist und eine große Erklärungskraft besitzt.
- Die Erklärungskraft ist dann groß, wenn die verbleibenden statistischen Schwankungen klein sind.
- Erklärungskraft und einfache Handhabung stehen i.A. im Gegensatz zueinander. Man kann z.B. immer eine Funktion finden, die durch alle Punkte des Streudiagramms läuft. Sie reduziert zwar die statistische Streuung auf Null, ist aber unsinnig kompliziert und hängt stark von der jeweiligen Stichprobe ab.

Lineare Regression

- Aus diesem Grunde beschränkt man die in Frage kommenden Funktionen von vornherein.
- Besonders einfach wird es, wenn man sich auf die Klasse der linearen Funktionen beschränkt:

$$f(X) = \alpha + \beta X$$

- Zu bestimmen sind nun die Parameter α und β, so dass die Erklärungskraft des Modells maximal wird und die statistische Streuung minimal.

Lineare Regression

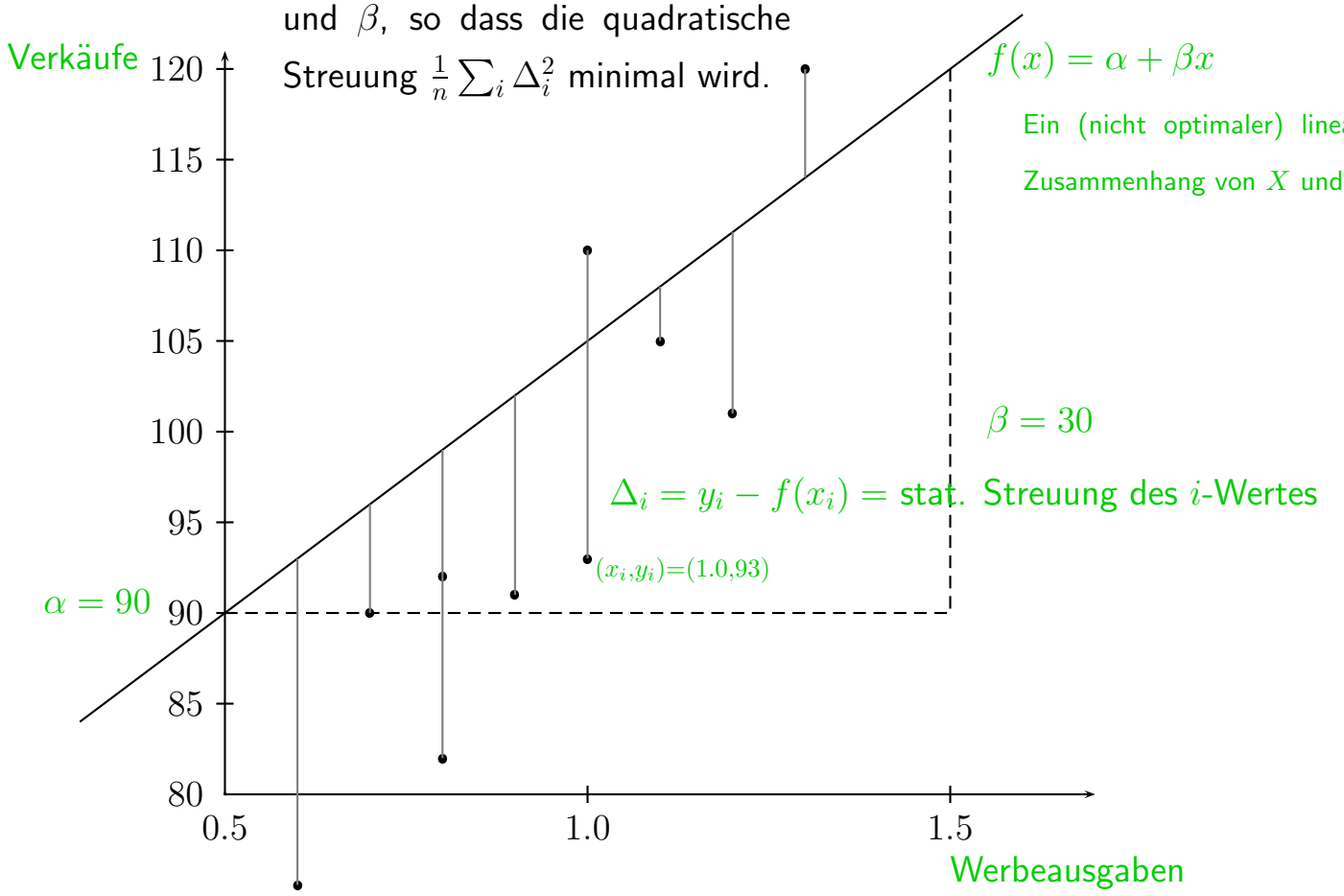
- Dazu brauchen wir noch ein Maß für die statistische Streuung. Wieder nehmen wir die mittlere quadratische Abweichung:

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

Definition

Die Gerade $f(x) = \hat{\alpha} + \hat{\beta} x$ mit den Parametern $\hat{\alpha}, \hat{\beta}$, für die die quadratische Abweichung minimal wird, heißt die Ausgleichsgerade zu den Wertepaaren aus der Urliste $(x_1, y_1), \ldots, (x_n, y_n)$.

Lineare Regression: Gesucht α und β, so dass die quadratische Streuung $\frac{1}{n}\sum_i \Delta_i^2$ minimal wird.



Lösung des Problems der linearen Regression:

Die Ausgleichsgerade $f(x) = \hat{\alpha} + \hat{\beta}x$ ist bestimmt durch

$$\hat{\beta} = \frac{c_{x,y}}{\sigma_x^2} \quad \text{und} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Beweis: Um das Minimum $(\hat{\alpha}, \hat{\beta})$ zu finden, müssen wir $Q(\alpha, \beta)$ nach α und β differenzieren und dann $= 0$ setzen:

$$\begin{aligned} 0 &\overset{!}{=} \\ 0 \overset{!}{=} \frac{\partial Q(\alpha,\beta)}{\partial \beta} &= \frac{2}{n}\sum_i (\alpha + \beta x_i - y_i)x_i = \frac{2}{n}\sum_i (\alpha x_i + \beta x_i^2 - x_i y_i) \end{aligned}$$

Die Lösung $(\hat{\alpha}, \hat{\beta})$ erfüllt also

$$\begin{aligned} 0 &= \hat{\alpha} + \hat{\beta}\bar{x} - \bar{y} \\ 0 &= \hat{\alpha}\bar{x} + \hat{\beta}\overline{x^2} - \overline{xy} = (\bar{y} - \hat{\beta}\bar{x})\bar{x} + \hat{\beta}\overline{x^2} - \overline{xy} = \hat{\beta}s_X^2 - c_{x,y} \end{aligned}$$