6 Korrelations- und Regressionsanalyse: Zusammenhangsanalyse stetiger Merkmale

6.1 Korrelationsanalyse

Jetzt betrachten wir bivariate Merkmale (X, Y), wobei sowohl X als auch Y stetig bzw. quasi-stetig und mindestens ordinalskaliert, typischerweise sogar intervallskaliert, sind. Am Rande wird auch der Fall gestreift, dass nur ein Merkmal quasi-stetig und das andere nominalskaliert ist.

6.1.1 Streudiagramm, Kovarianz- und Korrelationskoeffizienten

Bsp. 6.1.

- Nettomiete $\longleftrightarrow$ Wohnfläche
- Monatseinkommen $\longleftrightarrow$ Alter in Jahren
- Wochenarbeitseinkommen $\longleftrightarrow$ Wochenarbeitsstunden
- Wochenarbeitsstunden $\longleftrightarrow$ Hausarbeit in Stunden pro Woche
- Wochenarbeitsstunden (tatsächlich) $\longleftrightarrow$ Wochenarbeit (vertraglich)

6.1.2 Streudiagramme (Scatterplots)

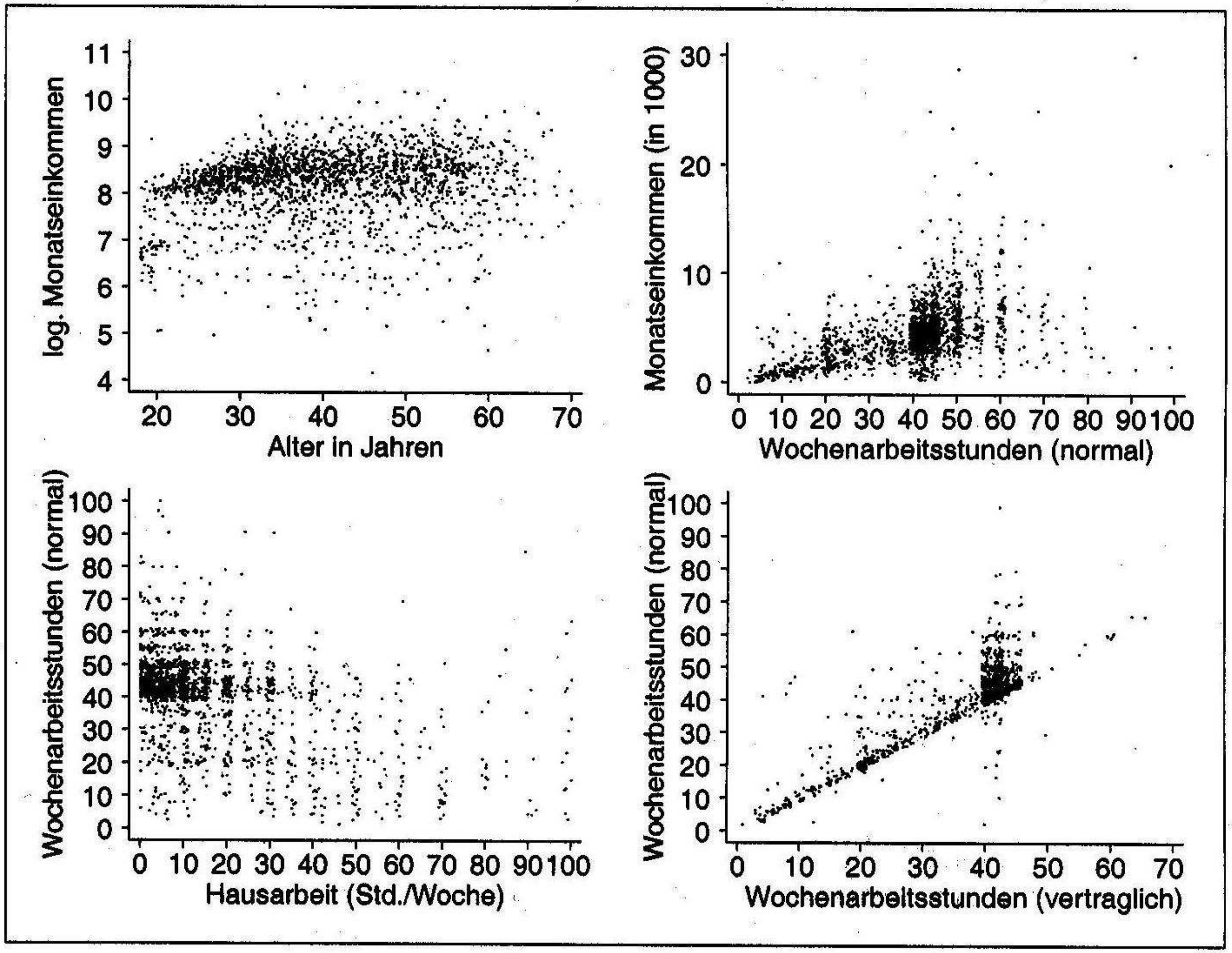
Sind die Merkmale stetig oder zumindestens quasi-stetig (sehr viele verschiedene Ausprägungen), werden Kontingenztabellen sehr unübersichtlich und praktisch aussagelos, da die einzelnen Häufigkeiten in den Zellen der Tabellen natürlicherweise durchwegs sehr klein sind.

Bessere Darstellungsform: *Scatterplot / Streudiagramm*:

Zeichne die Punkte (x_i, y_i), $i = 1, \ldots, n$, in ein X-Y-Koordinatensystem. Sind die Merkmale ordinal oder metrisch skaliert, so ergibt sich ein Eindruck der Struktur:

$\Longrightarrow$ Guter optischer Eindruck über das Vorliegen, die Richtung und gegebenenfalls die Art eines Zusammenhangs.

$\Longrightarrow$ Ausreißer werden leicht erkannt.



7

---

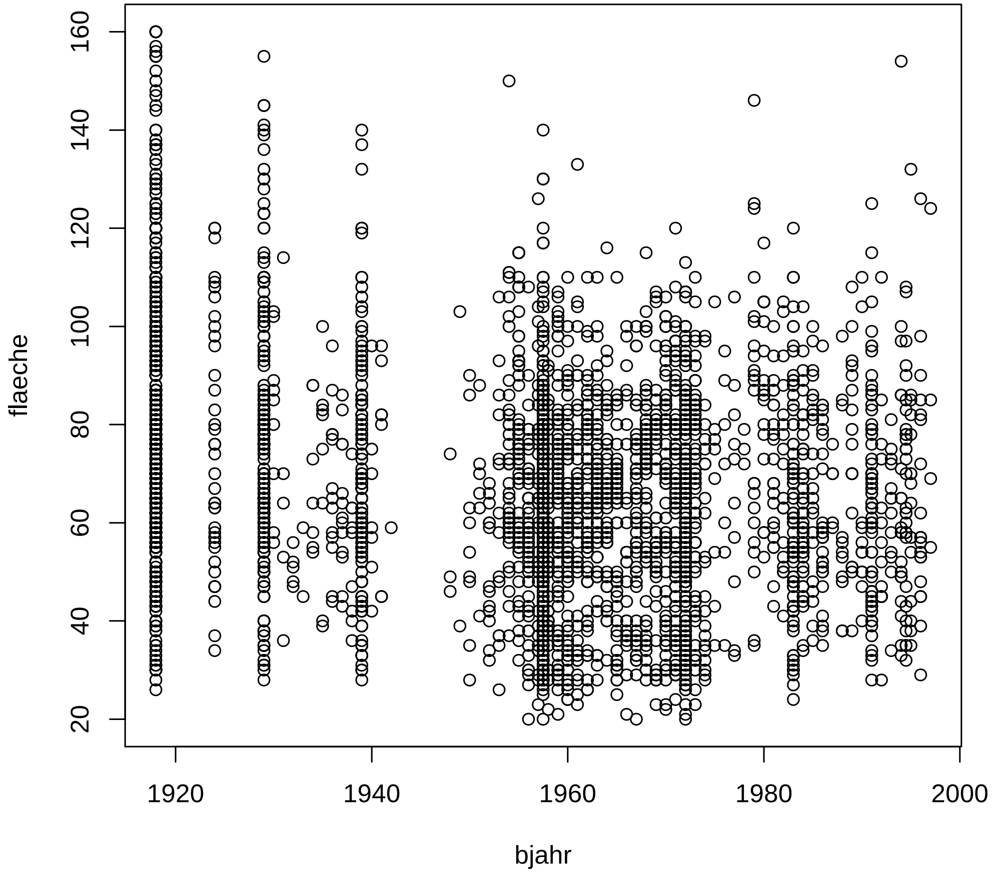
[7] Einige typische Scatterplots aus Jann (2002, p. 85 ff.)

Münchner Mietspiegeldaten (1999)
vgl. `www.regressionbook.org`

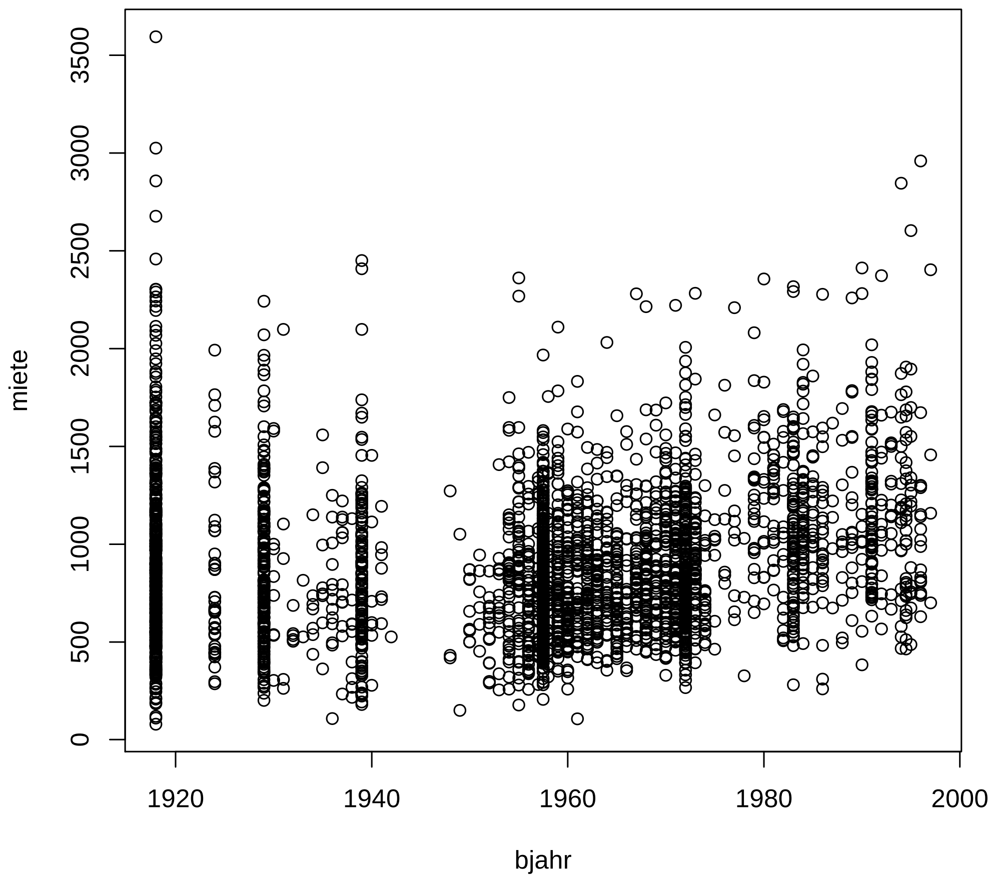
- Baujahr

- Wohnfläche

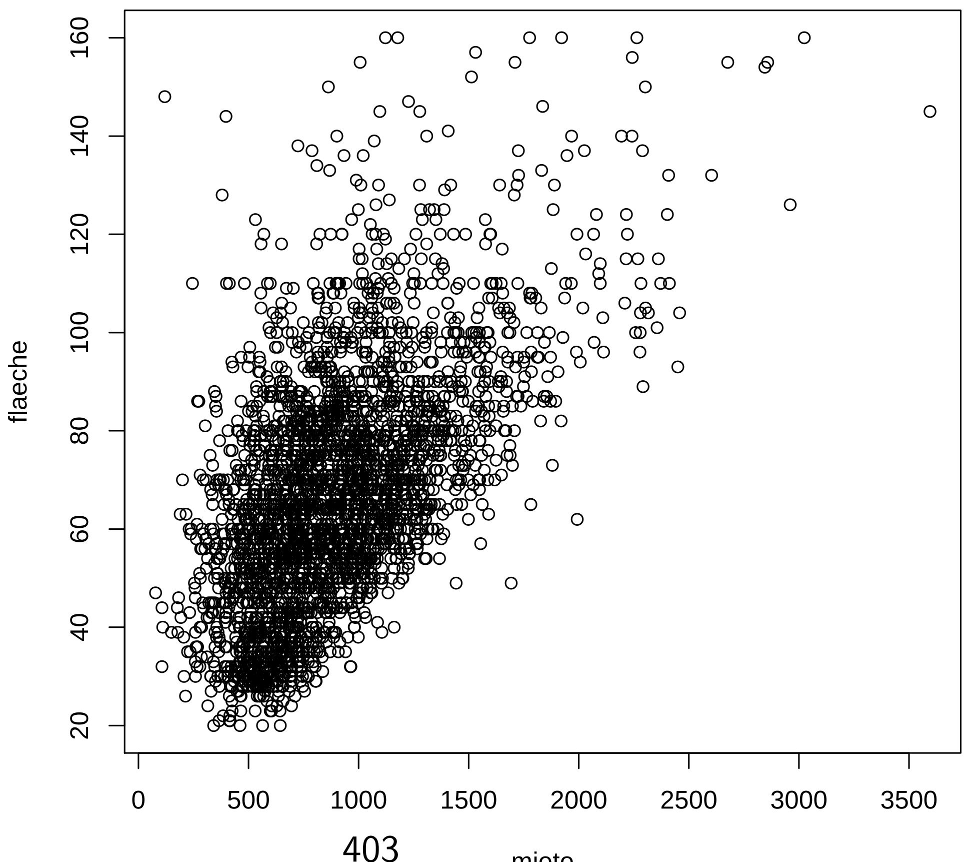
- monatliche Miete



flaeche
20
40
60
80
100
120
140
160
1920
1940
1960
1980
2000
bjahr



miete
3500
3000
2500
2000
1500
1000
500
0
1920
1940
1960
1980
2000
bjahr


flaeche
160
140
120
100
80
60
40
20
0
500
1000
1500
2000
2500
3000
3500
miete

6.1.3 Kovarianz und Korrelation

Wie misst man den Zusammenhang zwischen metrischen Merkmalen?

1. Eine Möglichkeit: betrachte wieder Paare $\{i^*, j^*\}$ von unterschiedlichen Einheiten und betrachte nicht nur **ob**, sondern auch, wie **stark** sie konkordant/diskordant sind. Dies kann beispielsweise über die Auswertung des Ausdrucks

   $$(x_{i^*} - x_{j^*})(y_{i^*} - y_{j^*})$$

   geschehen:

   - Das Vorzeichen gibt Auskunft über das Vorliegen von Konkordanz/Diskordanz: Bei positivem Vorzeichen müssen beide Faktoren des Produkts notwendigerweise das gleiche Vorzeichen haben. Dies bedeutet, dass Einheit i^* entweder in beiden Komponenten jeweils größer ist als Einheit j^*, oder aber in beiden Komponenten jeweils kleiner, was für eine Assoziation von großen (bzw. kleinen) Werten in X

mit großen (bzw. kleinen) Werten in Y spricht, also für Konkordanz.
Bei negativem Vorzeichen ist demgegenüber ein Faktor positiv und der andere negativ, was ein diskordantes Paar anzeigt.

- Die Größe des Betrages von $(x_{i^*} - x_{j^*})(y_{i^*} - y_{j^*})$ ist ein Maß für die Stärke der Konkordanz/Diskordanz des Paares, da der Ausdruck misst, wie weit die X- und Y-Werte des Paares auseinanderliegen. (Beachte, dass die Differenzen $x_{i^*} - x_{j^*}$ und $y_{i^*} - y_{j^*}$ für metrische Variablen sinnvoll interpretierbar sind).

2. Eine andere, sehr ähnliche Idee besteht darin, nach Konkordanz/Diskordanz zum **Schwerpunkt** zu fragen und ebenso auch die Abstände zur Messung der „individuellen Konkordanzstärke“ heranzuziehen. Dies führt im Wesentlichen zum gleichen Ergebnis, wie die erste Idee, nämlich zur sogenannten (empirischen) Kovarianz:

   - Betrachte den „Mittelpunkt“ der Daten $(\bar{x}, \bar{y})$ und dazu konkordante/diskordante Paare.

- Eine Beobachtung i mit Ausprägung (x_i, y_i) ist

  * $konkordant$ zu $(\bar{x}, \bar{y})$, spricht also für einen gleichgerichteten Zusammenhang, wenn

  $$(x_i > \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i < \bar{x} \text{ und } y_i < \bar{y}),$$

  also zusammengefasst wenn

  $$(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0.$$

  * $diskordant$ zu $(\bar{x}, \bar{y})$, spricht also für einen gegengerichteten Zusammenhang, wenn

  $$(x_i < \bar{x} \text{ und } y_i > \bar{y}) \text{ oder } (x_i > \bar{x} \text{ und } y_i < \bar{y}),$$

also zusammengefasst wenn

$$(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0.$$

- Wegen des metrischen Skalenniveaus sind auch die Abstände interpretierbar, das Produkt $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ gibt also sozusagen die Stärke der Konkordanz bzw. Diskordanz an.
- $(x_i - \bar{x})(y_i - \bar{y})$ ist positiv, wenn große (kleine) X-Werte mit großen (kleinen) Y-Werten einhergehen (gleichgerichteter Zusammenhang).
- $(x_i - \bar{x})(y_i - \bar{y})$ ist negativ, wenn große (kleine) X-Werte mit kleinen (großen) Y-Werten einhergehen (gegengerichteter Zusammenhang).

$\Longrightarrow$ Definiere als Zusammenhangsmaß die durchschnittliche individuelle Konkordanzstärke.

Definition: Gegeben sei ein bivariates Merkmal (X, Y) mit metrisch skalierten Variablen X und Y mit $\tilde{s}_X^2 > 0$ und $\tilde{s}_Y^2 > 0$. Dann heißen

$$\operatorname{Cov}(X, Y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

(empirische) Kovarianz von X und Y,

$$\varrho(X, Y) := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\tilde{s}_Y^2 \tilde{s}_X^2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

(empirischer) Korrelationskoeffizient nach Bravais und Pearson von X und Y, und

$$R_{XY}^2 := (\varrho(X, Y))^2 \tag{6.24}$$

Bestimmtheitsmaß von X und Y.

Bem. 6.2.

- Mit Idee 1) wären wir im Wesentlichen auch zur (empirischen) Kovarianz gekommen, es gilt:

$$\mathrm{Cov}(X,Y) = \frac{n-1}{n} \cdot \frac{1}{2} \cdot \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j} (x_i - x_j)(y_i - y_j)}_{\text{mittlere vorzeichenbehaftete Konkordanzstärke von Paaren}} \quad .$$

  Die (empirische) Kovarianz über die Konkordanz von Paaren auszurechnen ist aber umständlicher, da man alle Paare betrachten muss (vgl. Zusammenhangsmaße für ordinale Daten.)

- Die (empirische) Kovarianz $\mathrm{Cov}(X, Y)$ ist maßstabsabhängig.

- Wir werden später sehen: Das Teilen durch die Standardabweichungen normiert die Kovarianz und macht sie maßstabsunabhängig.

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sqrt{\tilde{s}_X^2}} \cdot \frac{(y_i - \bar{y})}{\sqrt{\tilde{s}_Y^2}} = \varrho(X, Y)$$

Also ist - im Sinne obiger Interpretation - der Korrelationskoeffizient die durchschnittliche *standardisierte* Konkordanzstärke.

- Die empirische Kovarianz ist eine Verallgemeinerung der empirischen Varianz. Die Kovarianz eines Merkmals mit sich selbst ist genau die empirische Varianz:

$$\begin{aligned}\mathrm{Cov}(X, X) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \tilde{s}_x^2\end{aligned}$$

- Anhand der Formel für die empirische Kovarianz sieht man auch, dass die Größe der Kovarianz für sich genommen unanschaulich zu interpretieren ist. Für den Korrelationskoeffizienten hingegen gilt:

$$-1 \leq \varrho(X, Y) \leq 1.$$

  und insbesondere $\varrho(X, X) = 1$.

- Viele der (un)angenehmen Eigenschaften der Varianz (z.B. Ausreißerempfindlichkeit) gelten in analoger Weise.

- Es gibt auch einen analogen Verschiebungssatz für die Kovarianz:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \ - \ \bar{x}\bar{y}$$

und damit

$$\varrho(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2}}.$$

Zur Erinnerung:

$$\tilde{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Bsp. 6.3.

Zunächst inhaltsleere Zahlenbeispiele, zur Interpretation später.

- Gegeben seien die Datenpaare

x_i	37	30	20	28	35
y_i	130	112	108	114	136

Tabelle:

	x_i^2	x_i	$x_i \cdot y_i$	y_i	y_i^2
		37		130	
		30		112	
		20		108	
		28		114	
		35		136	
$\sum$					

Es gilt: $\bar{x} = 30$ und $\bar{y} = 120$, sowie

$$\sum_{i=1}^{n} x_i^2 = 4678 \qquad \sum_{i=1}^{n} y_i^2 = 72600$$

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= 18282 \\ n &= 5 \end{aligned}$$

Basierend auf diesen Hilfsgrößen berechnet sich der Korrelationskoeffizient gemäß Verschiebungssatz als

$$\begin{aligned}\varrho(X,Y) &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2}} \\ &= \end{aligned}$$

- Gegeben sei ein Merkmal X und das Merkmal $Y = (X - 20)^2$ mit den Datenpaaren.

x_i	10	20	30
y_i	100	0	100

Bestimme die Kovarianz

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{3}(1000 + 0 + 3000) - 3 \cdot (\frac{10+20+30}{3}) \cdot (\frac{100+0+100}{3}) \\ &= \frac{4000}{3} - \frac{4000}{3} = 0 \end{aligned}$$

Für den Korrelationskoeffizienten ergibt sich damit ebenfalls $\varrho(X,Y) = 0$!

Was bedeutet das?

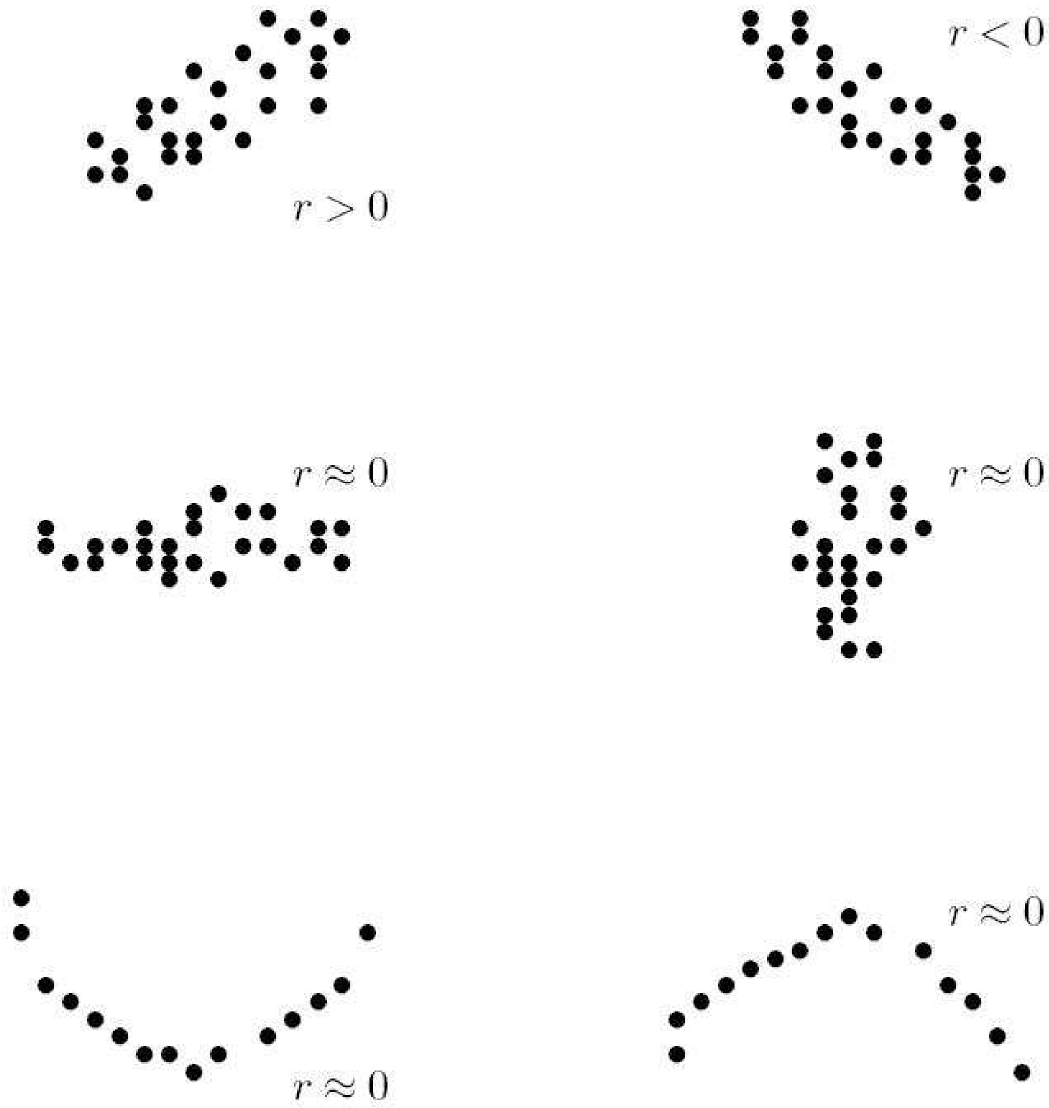
Beachte:

Bem. 6.4.

- Allgemein zeigt $|\varrho|$ und R^2 die Stärke eines *linearen* Zusammenhangs an, also wie gut sich die Datenpaare $(x_1, y_1), \ldots, (x_n, y_n)$ durch eine *Gerade* beschreiben lassen.
- Es gilt $|\varrho| = 1$ genau dann wenn es konstante Werte $a \neq 0$ und b gibt so, dass $Y = aX + b$, d.h. X und Y stehen in einem perfekten linearen Zusammenhang.
- Ist $\varrho = 0$ (und äquivalent dazu $\mathrm{Cov}(X, Y) = 0$), so nennt man X und Y *unkorreliert*. Es besteht dann keinerlei <u>linearer</u> Zusammenhang.
- Gelegentlich wird der Wert des Korrelationskoeffizienten darüberhinaus schematisch interpretiert (was aber meines Erachtens problematisch ist), z.B.
  - $|\varrho_{XY}| \leq 0.5$: schwache Korrelation.
  - $0.5 < |\varrho_{XY}| \leq 0.8$: mittlere Korrelation.
  - $|\varrho_{XY}| > 0.8$: starke Korrelation.
- R^2 ist ein PRE-Maß, das misst, welchen Anteil der gesamten Variation sich durch

einen linearen Zusammenhang beschreiben lässt. (Näheres dazu im Abschnitt über die Regression.)

- Die Betonung der *Linearität* des Zusammenhangs ist wesentlich.


$r > 0$
$r < 0$
$r \approx 0$
$r \approx 0$
$r \approx 0$
$r \approx 0$

- Die Zusammenhangsmaße sind invariant gegenüber Vertauschen von Y und X, unterscheiden also nicht welche Variable als abhängige, welche als unabhängige gilt:

$$\varrho(X,Y) = \varrho(Y,X) \qquad R_{XY} = R_{YX}.$$

- Im Gegensatz zur Kovarianz sind $\varrho(X,Y)$ und R^2_{XY} invariant gegenüber streng monoton steigenden linearen Transformationen. Genauer gilt mit $\tilde{X} := a \cdot X + b$ und $\tilde{Y} := c \cdot Y + d$

$$\varrho(\tilde{X},\tilde{Y}) = \varrho(X,Y)$$

falls $a \cdot c > 0$ und

$$\varrho(\tilde{X},\tilde{Y}) = -\varrho(X,Y)$$

falls $a \cdot c < 0$. Die Korrelation ist also in der Tat maßstabsunabhängig.

Bsp. 6.5.

(aus Jann (2002) S.87ff)

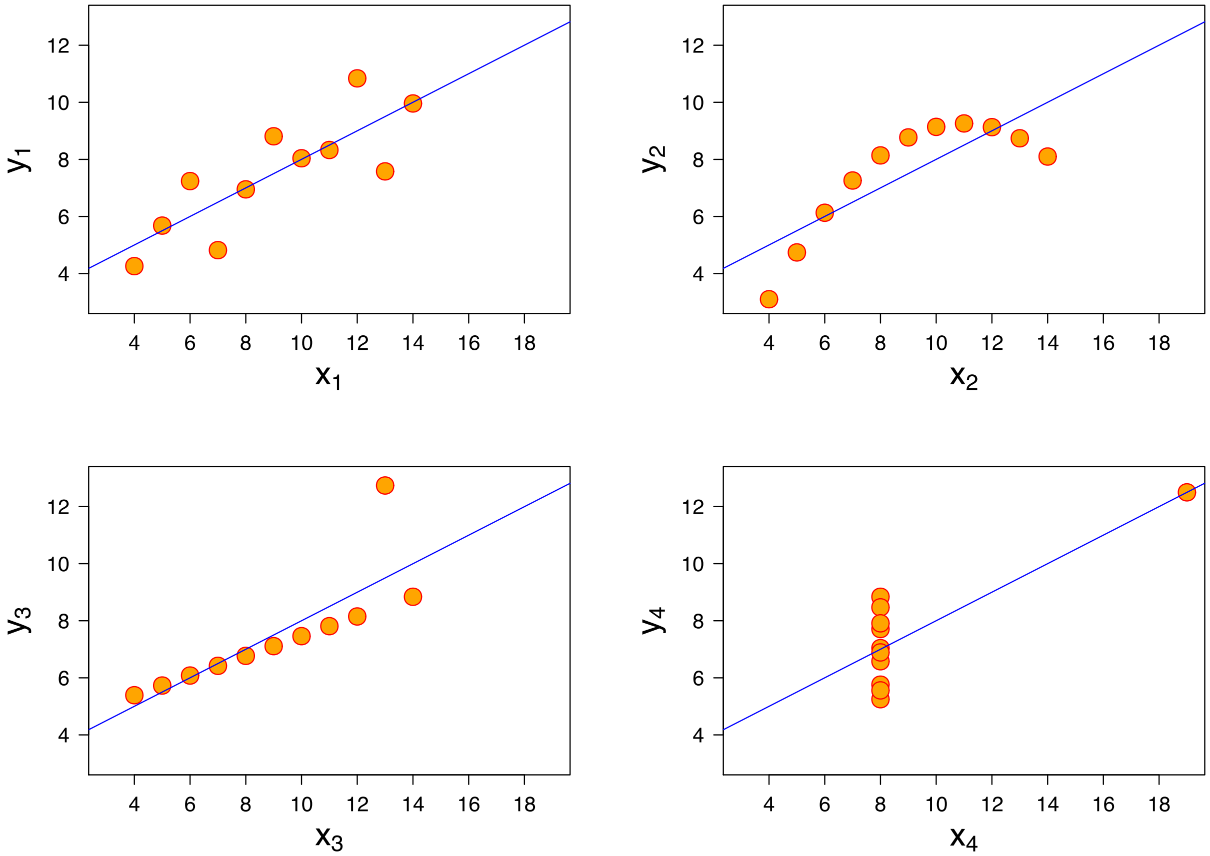
- Arbeitsstunden und Erwerbseinkommen: 0.495
  moderater positiver Zusammenhang.
- Arbeitsstunden und Haushalt: -0.434
  moderater negativer Zusammenhang.
- Vertragliche und geleistete Wochenarbeitsstunden: 0.868
  hoch positiv korreliert (Punkte liegen sehr nahe an „bester Gerade“).

Ergebnisse bei Mietspiegeldaten:
Korrelation nach Pearson

- (Fläche, Miete) 0.5845306
- (Baujahr, Miete) 0.1341479

- (Baujahr, Fläche) -0.2306214

Bsp. 6.6. *Anscombe's Quartet*



(Quelle: Wikipedia; Anscombe's quartet)