

WIRTSCHAFTSSTATISTIK

MODUL 6: KORRELATION UND REGRESSION

WS 2020/21

DR. E. MERINS

ZUR GEMEINSAMEN ANALYSE MEHRERER MERKMALE

- Bei einer empirischen Untersuchung werden in der Regel **mehrere Merkmale** gemessen, z.B. X, Y, Z , usw.
- Merkmale werden einzeln ausgewertet (Charakterisierung und Bereinigung) → **univariate Analyse**
- Mehrere Merkmale werden gemeinsam ausgewertet → **multivariate Analyse**

$$X \leftrightarrow Y$$

Zusammenhang (z. B. : Lohnniveau \leftrightarrow Preisniveau)

$$X \rightarrow Y$$

Abhängigkeit (z. B. : Preis \rightarrow Absatzmenge)

ZUR GEMEINSAMEN ANALYSE MEHRERER MERKMALE

- Die **Analyse von Zusammenhängen** ist ein Teilgebiet der **multivariaten** (lat.: multus - vielfach, varia - Allerlei) Statistik. Dabei erfassen statistische Untersuchungen mehrere Merkmale eines Merkmalsträgers gleichzeitig (sogenannte multivariate Datensätze).
- Bei der Analyse von Zusammenhängen können folgende Fragestellungen auftreten:
Besteht überhaupt ein Zusammenhang zwischen Merkmalen (oder Variablen)?
Und wenn ja, dann:
 - Wie stark ist dieser Zusammenhang?
 - Wie lässt sich die Stärke (der Grad, die Intensität) des Zusammenhangs bzw. die Abhängigkeit zwischen zwei (oder mehreren) Merkmalen messen?
 - Lässt sich der Zusammenhang in einer bestimmten Form (Typ, Art) darstellen?
 - Lassen sich die beobachteten Werte einer Variable X durch die Werte einer oder mehrerer anderen Variablen Y (Y_1, Y_2, \dots) näherungsweise bestimmen?

ZUSAMMENHANGSANALYSE

- **Korrelationsanalyse** (oder Maßkorrelationsanalyse) → wird geprüft, ob zwei Variablen X und Y **linear zusammenhängen** und **wie stark** dieser **Zusammenhang** ist
- Korrelationsanalyse mit dem Spezialfall der **Rangkorrelationsanalyse** → **Zusammenhang zweier ordinalskalierter Merkmale** mit Hilfe von Rangzahlen. Der **Rangkorrelationskoeffizient nach SPEARMAN** hat eine besondere praktische Bedeutung wegen seiner einfachen Berechnung
- **Kontingenzanalyse** (oder Assoziationsanalyse, lat.: contingentia - Zufälligkeit) → **Zusammenhangsanalyse** auf der Basis einer Kontingenztafel (=Häufigkeitstabelle, s. Modul 03). Je größer der Unterschied zwischen den Häufigkeiten in den Tabellenfeldern ist, umso stärker ist der Zusammenhang bzw. die Abhängigkeit zwischen den Merkmalen

Hinweis: Typen und Arten des Zusammenhanges sind in den Kurs-Materialien „**Abschnitt IV.**

Multivariate Daten Teil 10: ZHA-Zusammenhänge“ gut beschrieben

ZUSAMMENHANGSANALYSE

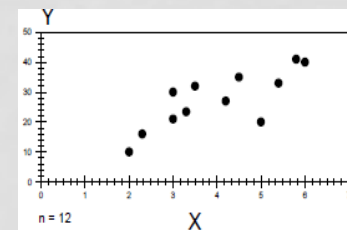
X, Y, Z Merkmale

Zusammenhangsanalyse (Interdependenzanalyse)

Es wird eine Wechselwirkung der Variablen untereinander untersucht. Ein Zusammenhangsmaß oder auch Assoziationsmaß genannt, gibt in der Statistik die Stärke und gegebenenfalls die Richtung eines Zusammenhangs

Zusammenhangsanalyse zwischen zwei metrischen Merkmalen X und Y

- Zusammenhangsanalyse → Korrelationsanalyse
- Zusammenhangsmaß → Korrelationskoeffizient $-1 \leq r_{xy} \leq +1$
- Grafische Darstellung → Streudiagramm



ABHÄNGIGKEITSANALYSE

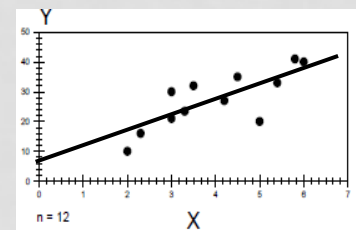
X, Y, Z Merkmale

Abhängigkeitsanalyse (Dependenzanalyse)

Es wird zwischen unabhängigen und abhängigen Merkmalen unterschieden. Es geht um einen gerichteten Zusammenhang. Der Anwender hat vorab eine sachlogisch begründete Vorstellung über den Zusammenhang zwischen den Merkmalen, d.h. er weiß oder vermutet, welche der Merkmale auf andere Merkmale einwirken (können)

Abhängigkeitsanalyse zwischen zwei metrischen Merkmalen X und Y

- Abhängigkeits**analyse** → **Regressionsanalyse**
- Abhängigkeits**maß** → **Regressionsfunktion** $\hat{y} = a + b \cdot x$
- **Grafische** Darstellung → **Streudiagramm**
+ **Regressionsgerade**



MULTIVARIATE ANALYSEMETHODEN

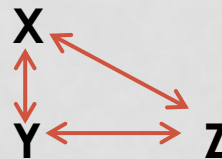
X, Y, Z Merkmale

Beispiel:

Zusammenhangsanalyse (Interdependenzanalyse)

Mitarbeiterzufriedenheit

Kundenzufriedenheit

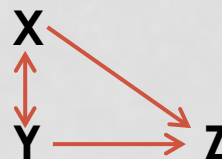


Motivation der Mitarbeiter

Abhängigkeitsanalyse (Dependenzanalyse)

Verkaufsfläche

Anzahl Personal



Filialumsatz

ZUSAMMENHANGSANALYSE BEI NICHT METRISCHEN MERKMALEN

Rangkorrelationskoeffizient

- ein Maß für die Stärke des Zusammenhangs zweier ordinalskaliertter Merkmale
- der Spearmansche Rangkorrelationskoeffizient nutzt Ränge statt der Beobachtungswerte → ein Spezialfall von Pearsons Korrelationskoeffizient, bei dem die Daten in Ränge konvertiert werden, bevor der Korrelationskoeffizient berechnet wird
- benötigt keine Annahme, dass die Beziehung zwischen den Variablen linear ist
- robust gegenüber Ausreißern

Kontingenzkoeffizient

- ein Maß für die Stärke des Zusammenhangs zweier (oder mehrerer) nominaler oder ordinaler Merkmale. Er basiert auf dem Vergleich von tatsächlich ermittelten Häufigkeiten zweier Merkmale mit den Häufigkeiten, die man bei Unabhängigkeit dieser Merkmale erwartet hätte
- kann bei beliebig großen Kreuztabellen angewendet werden
- der Kontingenzkoeffizient C liegt zwischen 0 und +1, d.h., $0 \leq C \leq 1$.

Phi-Koeffizient (auch Vierfelder-Korrelationskoeffizient)

- ein Maß für die Stärke des Zusammenhangs zweier dichotomer Merkmale
- basiert auf einer Kontingenztafel, die die gemeinsame Häufigkeitsverteilung der Merkmale enthält

STREUDIAGRAMM

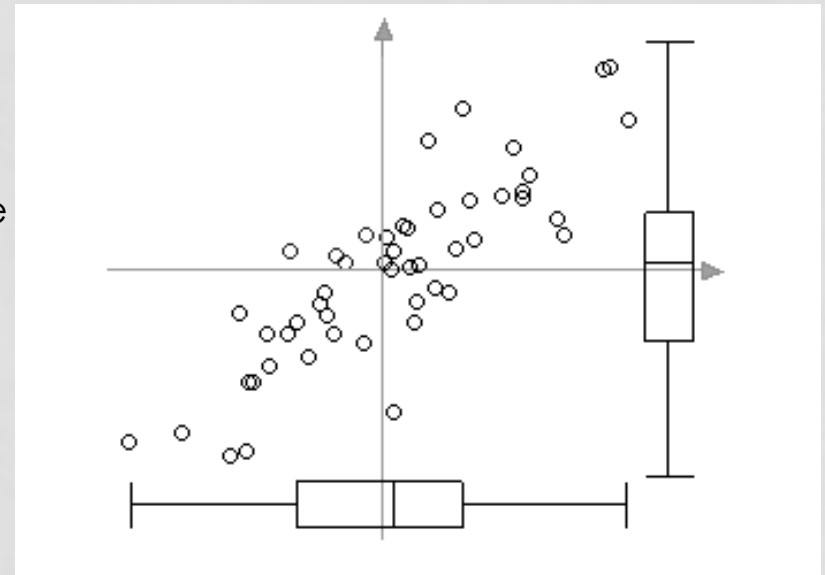
Streudiagramm (oder Streuungsdiagramm)

Ein Streudiagramm (*engl. scatter plot*) ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt.

Bei beiden Boxplots stimmt der eingetragene Median fast mit der Koordinatenachse überein, es gibt also jeweils etwa gleich viele positive und negative Werte, gemeinsam nehmen die Variablen aber fast nur Werte im I. und im III. Quadranten ein.

Aus Lage und Form der dargestellten Punktwolke lassen sich die **Stärke** und die **Richtung** des Zusammenhangs der Merkmale ablesen.

Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen.



KORRELATION

- **Korrelation** → zahlenmäßiger statistischer Zusammenhang zwischen zwei Merkmalen X und Y.

Eine **positive Korrelation** liegt vor, wenn die beiden Merkmale sich gleichförmig entwickeln → bei höheren Werten von X auch Y hohe Werte hat.

Eine **negative Korrelation** liegt vor, wenn X und Y sich gegenläufig entwickeln → bei höheren Werten von X liegen niedrigere Werte von Y vor.

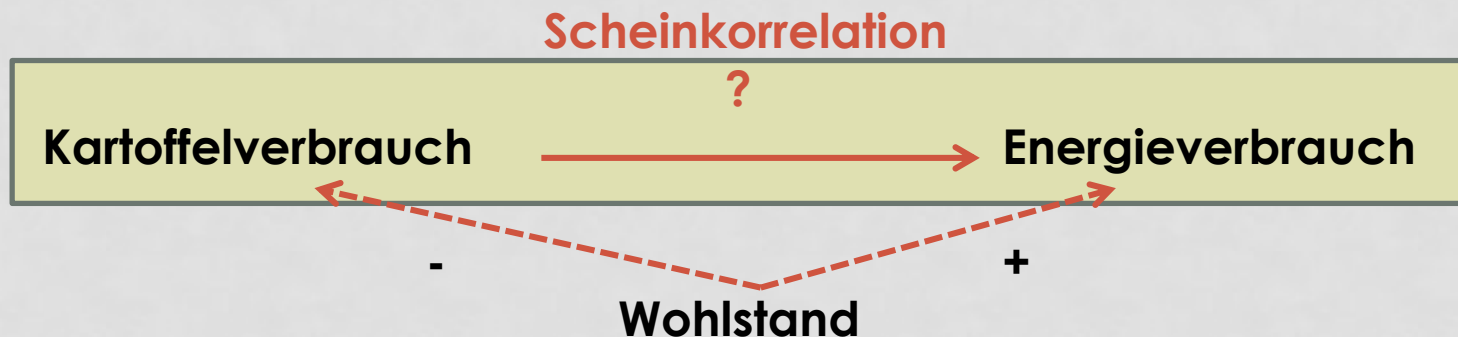
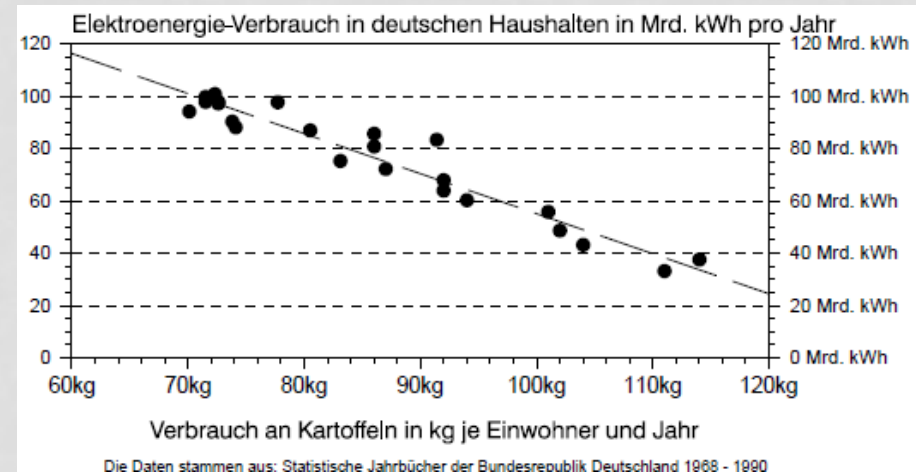
- Ein **kausaler Zusammenhang** zwischen X und Y liegt vor, wenn es zwischen X und Y eine Ursache-Wirkungs-Beziehung gibt, d.h., wenn eine Veränderung des abhängigen Merkmals Y eindeutig auf eine Veränderung von X zurückzuführen ist.
- Eine Korrelation sagt nichts über einen kausalen Zusammenhang aus und auch nichts über eine Kausalitätsrichtung.

PROBLEME BEI DER ABHÄNGIGKEITSANALYSE

Problem: **Abhängigkeitsanalyse muss sinnvoll sein!!**
(Korrelation \neq Kausalität)

Kausalität → eine Ursache-Wirkungs-Beziehung zw. X und Y, d.h. wenn eine Veränderung des einen Merkmals eine Veränderung bei dem anderen Merkmal hervorruft.

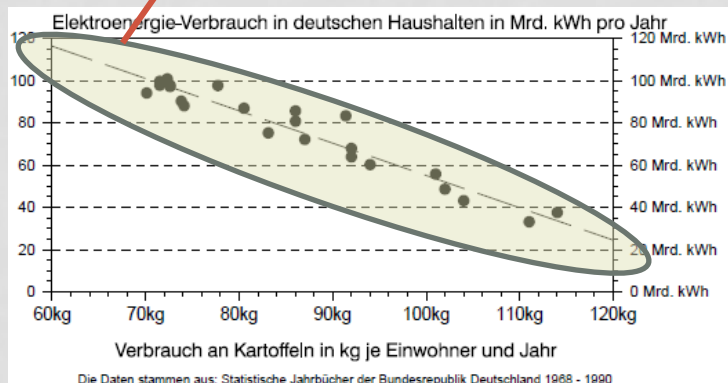
Korrelation → ist eine notwendige aber keine hinreichende Voraussetzung für einen kausalen Zusammenhang. Quantifizierung erfolgt über den **Korrelationskoeffizienten**.



SCHEINKORRELATION

Korrelation
statistischer Zusammenhang

Energie-Sparen durch höheren
Kartoffel-Verbrauch ?

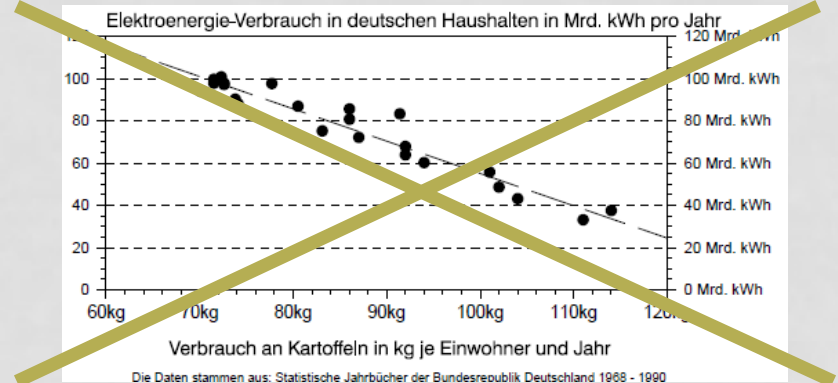


Kausalität

kausaler Zusammenhang =
Ursache-Wirkungs-Beziehung
... denn sonst könnte man den
Energie-Verbrauch durch Kartoffel-
Verbrauch beeinflussen ...

Scheinkorrelation
kein kausaler Zusammenhang

Energie-Sparen durch höheren
Kartoffel-Verbrauch ?



SCHEINKORRELATION

Das wohl berühmteste Beispiel für eine Scheinkorrelation: Der Storch bringt die Babys!

Der Wissenschaftler Robert Matthews fand 2001 eine **Korrelation** nicht unerheblicher Höhe von 0,62 zwischen der Geburtenrate eines Landes und der Anzahl dort lebenden Störche.

Wie kommt diese Korrelation zustande?

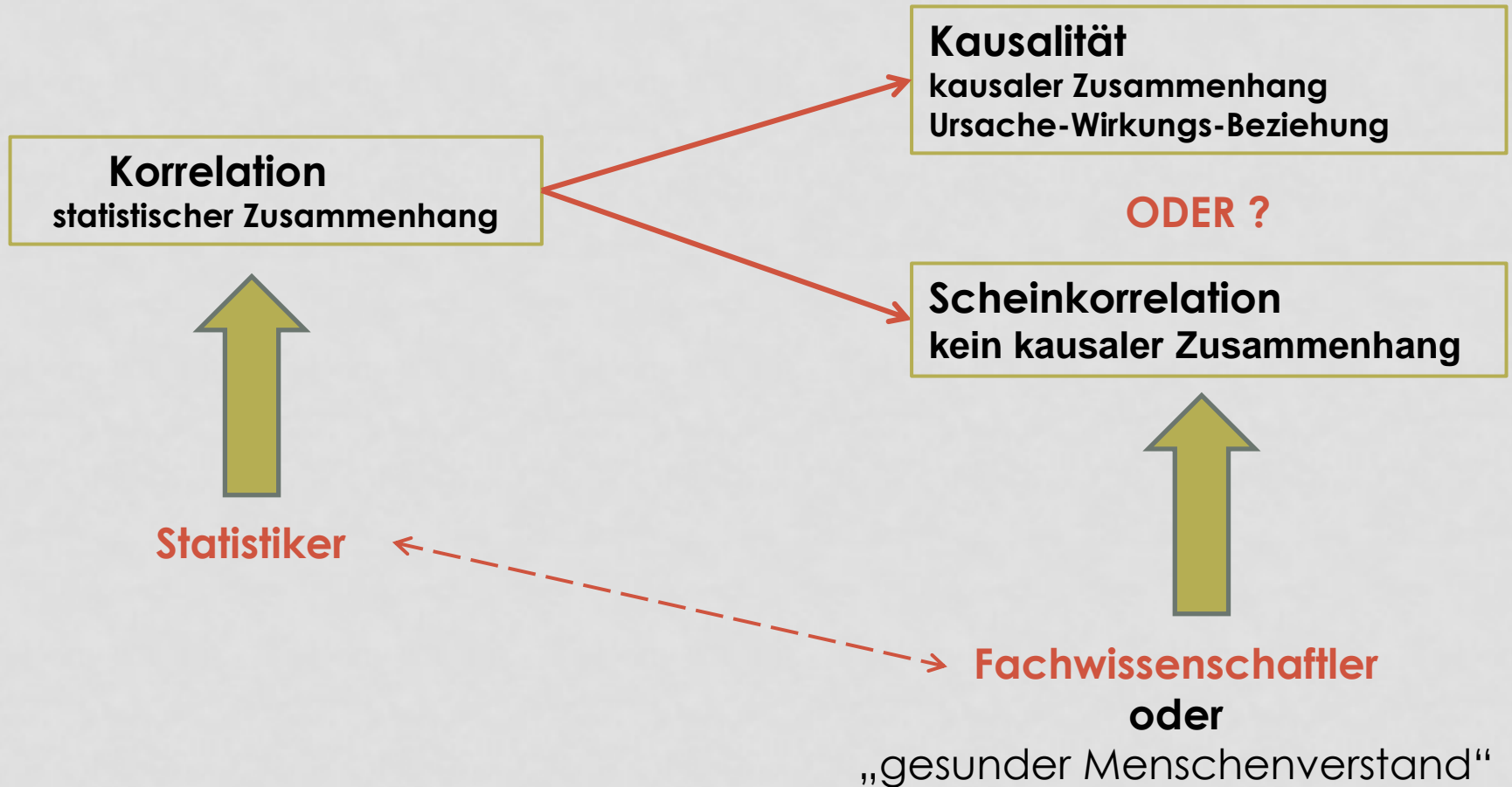
Bei Matthews basiert diese hohe Korrelation zu einem großen Teil auf der Größe des Landes: in größeren Ländern leben mehr Störche. Und dort werden mehr Kinder geboren als in kleineren Ländern. Auch eine andere Erklärung wäre denkbar → „**Urbanität vs. Ländlichkeit**“. In der Stadt leben weniger Störche als auf dem Land. Gleichzeitig ist auf dem Land aufgrund soziokultureller Unterschiede die Geburtenrate höher als in der Stadt. Daraus ergibt sich, dass in Gegenden, in denen viele Störche leben, auch die Geburtenrate höher ist.

Auf jeden Fall:

Ein **kausaler Zusammenhang** liegt nicht vor, der Storch bringt keine Kinder!



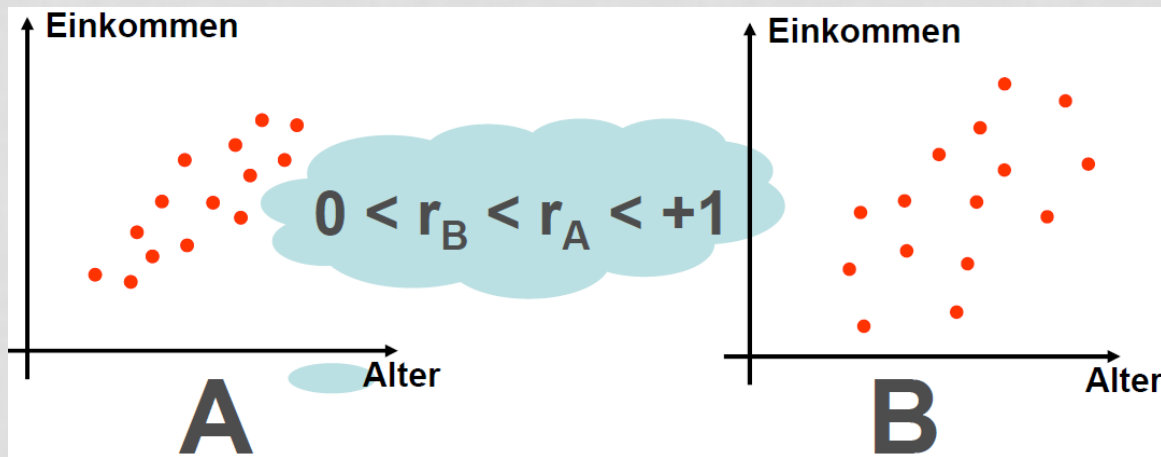
KORRELATION



KORRELATIONSKOEFFIZIENT

Korrelationskoeffizient r_{xy}

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



Korrelationskoeffizient → statistische Kennzahl, die informiert über

- die Stärke des linearen Zusammenhangs
- die Richtung des linearen Zusammenhangs

→ **Korrelationskoeffizient** ist ein dimensionsloses **Maß für den Grad des linearen Zusammenhangs**

KORRELATIONSKOEFFIZIENT

Korrelationskoeffizient r_{xy}

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

Positiver Zusammenhang $r > 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit hohen Werten in der anderen Variablen auf.

Negativer Zusammenhang $r < 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit niedrigen Werten in der anderen Variablen auf.

Korrelationskoeffizient $r = -1$: es liegt ein extrem starker negativer linearer Zusammenhang vor → die Punktwolke liegt auf einer Geraden mit negativer Steigung.

Korrelationskoeffizient $r = +1$: es liegt ein extrem starker positiver linearer Zusammenhang vor → die Punktwolke liegt auf einer Geraden mit positiver Steigung.

Korrelationskoeffizient $r = 0$: es liegt kein linearer Zusammenhang vor.

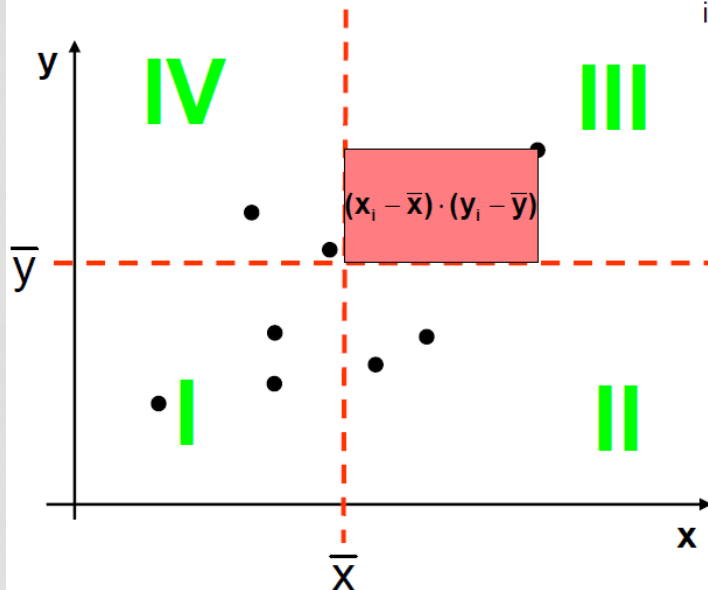
KORRELATIONSRECHNUNG (NACH PEARSON)

Definition:

Für zwei mindestens intervallskalierten Merkmale X und Y mit jeweils positiver Standardabweichung und Kovarianz ist der Korrelationskoeffizient (Pearsonscher Maßkorrelationskoeffizient) definiert durch:

$$r_{XY} = \frac{\text{COV}(X, Y)}{s_X \cdot s_Y}$$

Kovarianz $\text{COV}(X, Y) = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$



(x_i, y_i) in Quadrant	Vorzeichen von $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
I	+
II	-
III	+
IV	-

Die Kovarianz $\text{COV}(X, Y)$ informiert über die gemeinsame Variabilität der beiden Merkmale X und Y.

Ist der Zusammenhang positiv, dann ist die Kovarianz positiv, ist der Zusammenhang negativ, dann ist die Kovarianz negativ. Gibt es keinen (linearen) Zusammenhang zwischen X und Y, dann liegt die Kovarianz in der Nähe von 0.

KORRELATIONSRECHNUNG (NACH PEARSON)

„Standardisierung“ der Kovarianz:

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

BEISPIEL

Beispiel:

Verkaufsfläche → Filialumsatz

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)
1	3	30
2	2	10
3	6	40
4	5	20
Summe	16	100

KORRELATIONSRECHNUNG

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$r_{XY} = r = \frac{\text{COV}(X, Y)}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y} = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i \cdot y_i \right) - \bar{x} \cdot \bar{y}}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2} \cdot \sqrt{\left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2}}$$

KORRELATIONSRECHNUNG

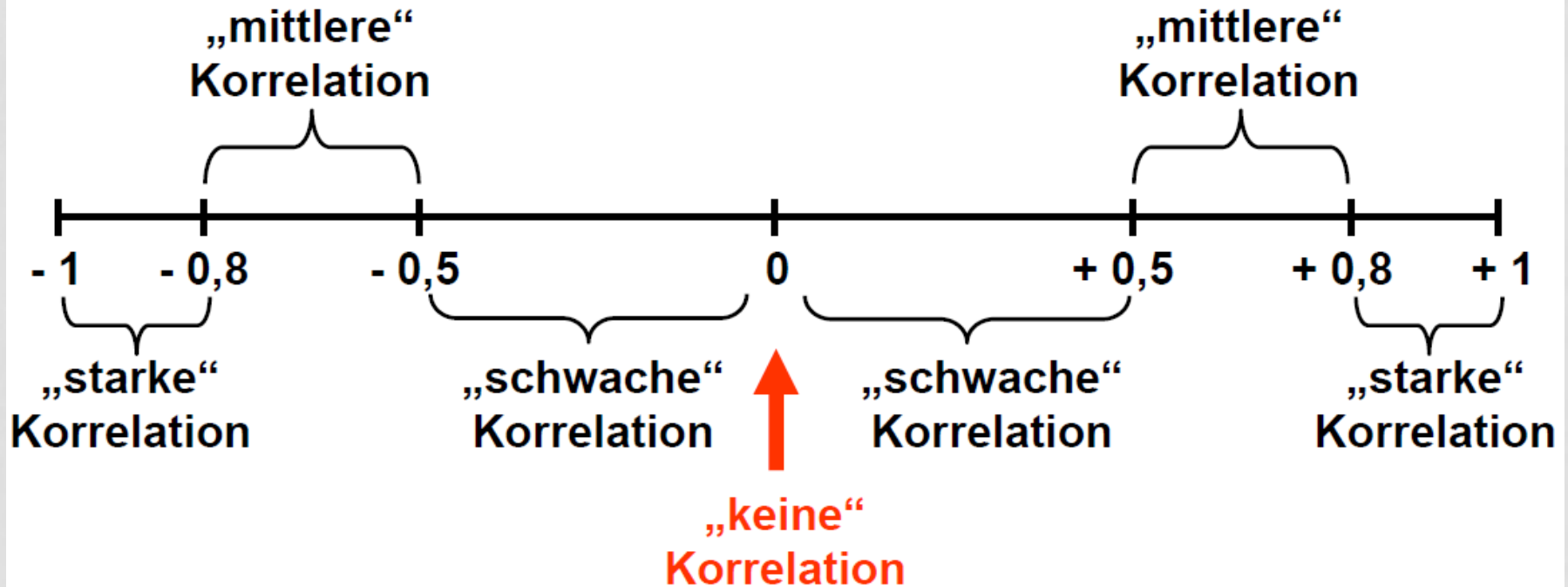
Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

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

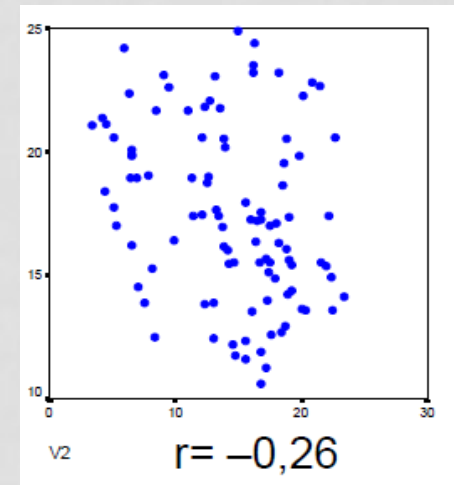
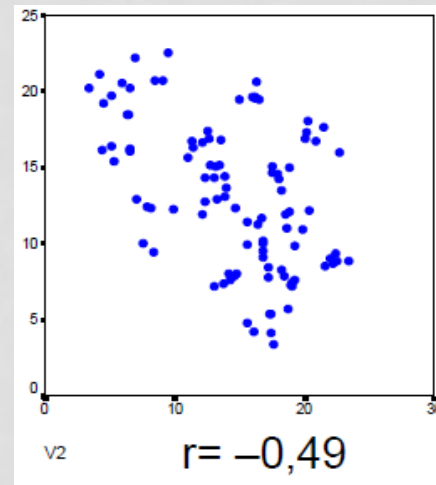
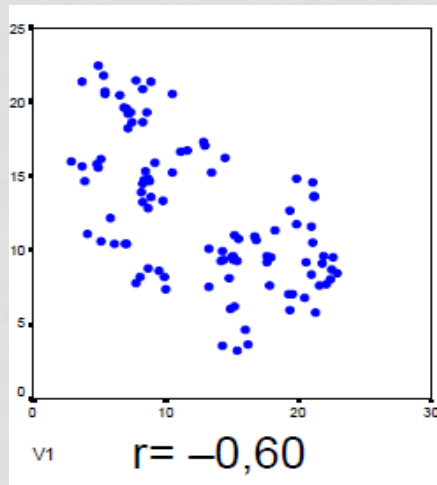
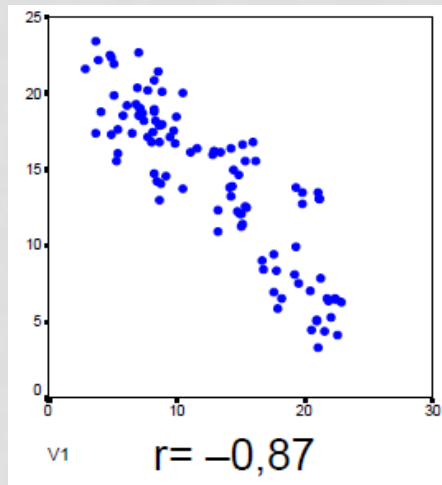
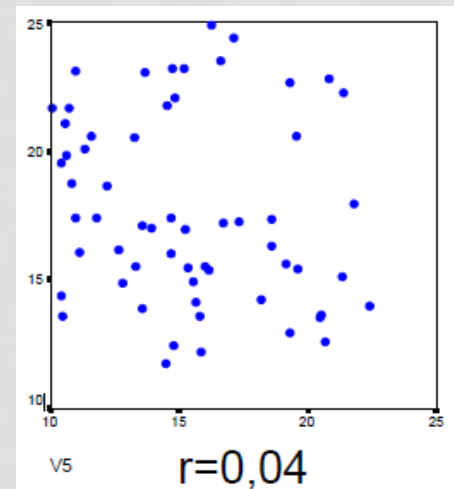
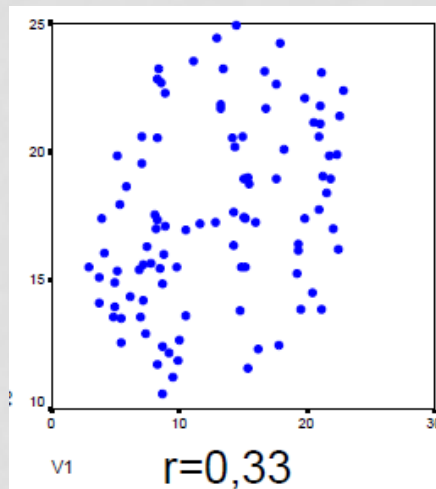
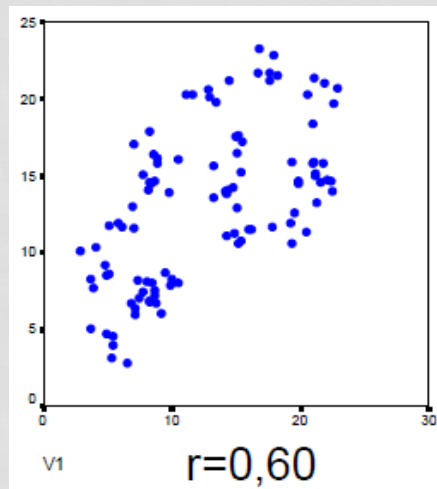
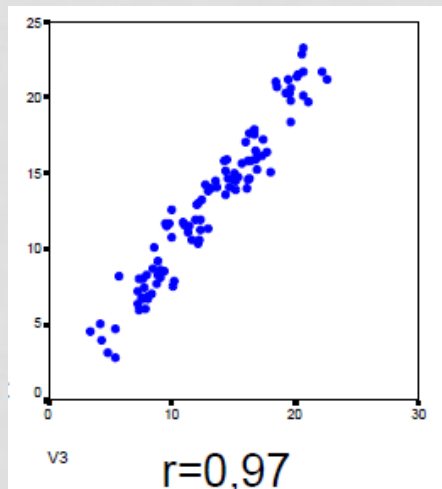
$$\frac{\frac{1}{4} \cdot 450 - 4 \cdot 25}{\sqrt{\frac{1}{4} \cdot 74 - 4^2} \cdot \sqrt{\frac{1}{4} \cdot 3000 - 25^2}} = \frac{112,5 - 100}{\sqrt{2,5} \cdot \sqrt{125}} = \frac{12,5}{1,581 \cdot 11,180} = \frac{12,5}{17,676} = +0,707$$

ZUR INTERPRETATION DES KORRELATIONSKOEFFIZIENTEN



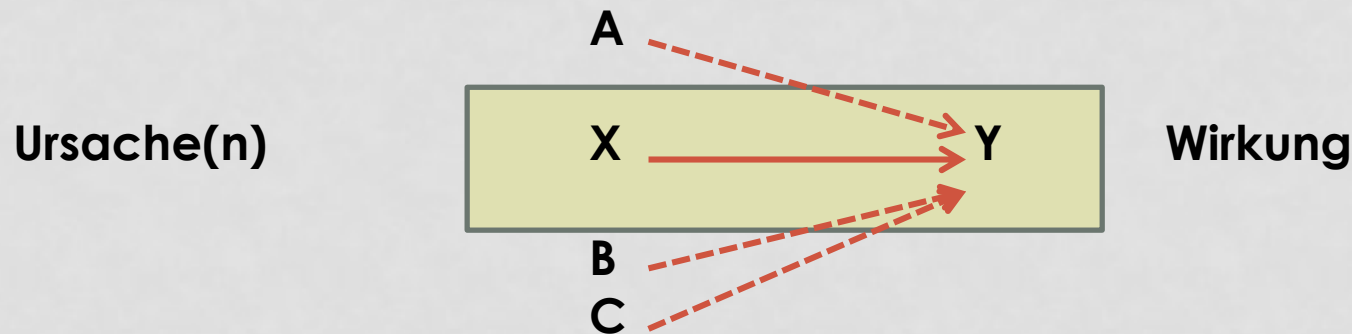
Nur Orientierungswerte → Entscheidung ist immer problembezogen!!!

ZUR INTERPRETATION DES KORRELATIONSKOEFFIZIENTEN



PROBLEME BEI DER REGRESSIONSANALYSE

Problem: Es gibt meist nicht nur einen Einflussfaktor
(*Probleme sind selten monokausal ...*)



- **einfache** Regressionsanalyse

→ zwei metrischen Größen: Einflussgröße X und Zielgröße Y. Es wird mithilfe von zwei Parametern eine Gerade durch eine Punktwolke gelegt, sodass der lineare Zusammenhang zwischen X und Y möglichst gut beschrieben wird.

- **multiple** Regressionsanalyse

→ eine Verallgemeinerung der einfachen linearen Regression mit k Regressoren, welche die abhängige Variable erklären sollen.

Mehrere metrischen Größen: mehrere Einflussgrößen X_1, \dots, X_k und eine Zielgröße Y.

REGRESSIONSFUNKTION

Voraussetzungen:

- X und Y quantitative (metrische) Merkmale
- $X \rightarrow Y$ (es existiert ein Zusammenhang)

Vorbereitende Arbeiten:

- Überprüfung, ob Abhängigkeitsanalyse sinnvoll ist
- Erhebung von Daten für X und Y $\rightarrow (x_1, y_1) , \dots , (x_n, y_n)$

1. Schritt: Visualisierung im Streudiagramm
(qualitative Abhängigkeitsanalyse)

2. Schritt: Auswahl eines Funktionstyps
(hier: Beschränkung auf lineare Funktionen)

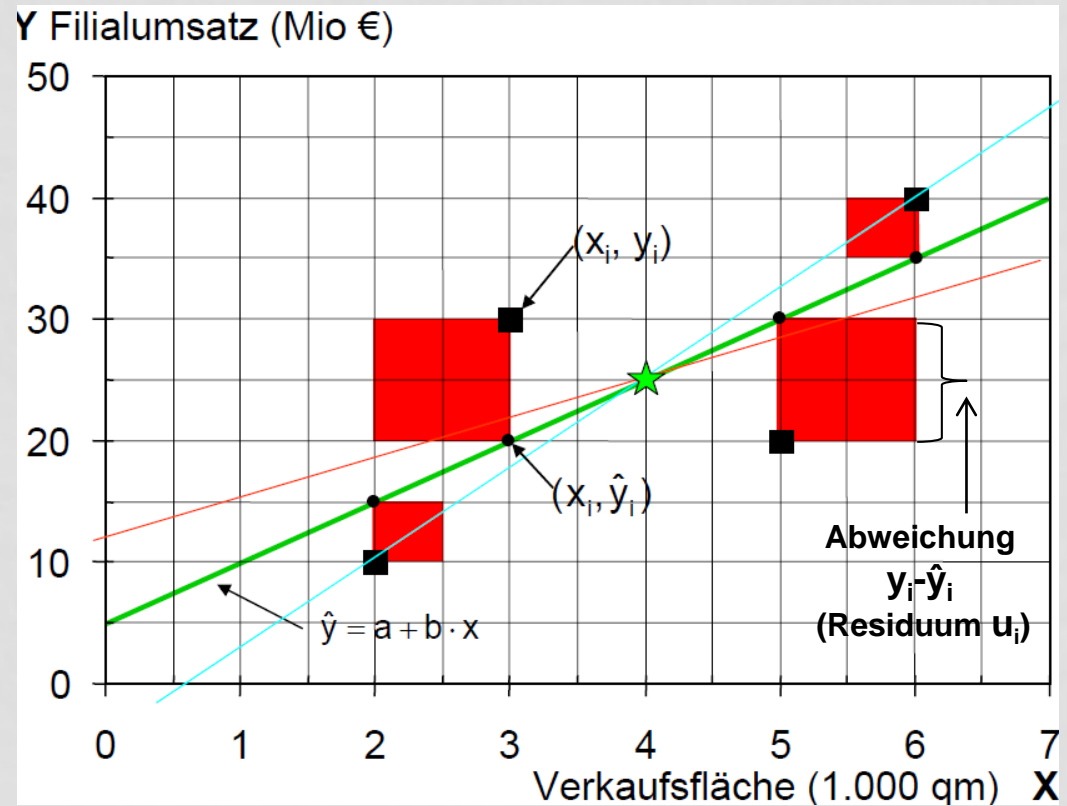
3. Schritt: Berechnung der Regressionsfunktion
(nach Methode der kleinsten Quadrate)

REGRESSIONSFUNKTION

Beispiel:

Verkaufsfläche → Filialumsatz

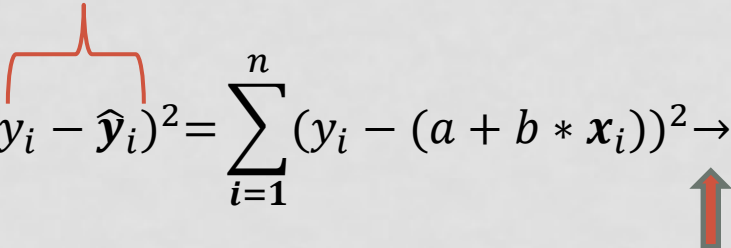
Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)
1	3	30
2	2	10
3	6	40
4	5	20
Summe	16	100



REGRESSIONSFUNKTION

Bestimmung der **Regressionsfunktion** $\hat{y} = a + b \cdot x$ nach der **Methode der kleinsten Quadrate**:

Die **Regressionskoeffizienten** **a** und **b** (**Kurvenparameter**) werden so bestimmt, dass die Summe der quadratischen Abweichungen der Kurve von den beobachteten Punkten minimal ist:

$$\text{Residuen } u_i$$
$$OLS(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 \rightarrow \min$$


Methode der kleinsten Quadrate (OLS = ordinary least squares)

Es wird die (partielle) Differentialrechnung genutzt, um die Extremwertaufgabe „Minimierung der Summe der Abweichungsquadrate“ zu lösen.

REGRESSIONSRECHNUNG

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$a = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{74 \cdot 100 - 16 \cdot 450}{4 \cdot 74 - 16^2} = \frac{7400 - 7200}{296 - 256} = \frac{200}{40} = 5$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{4 \cdot 450 - 16 \cdot 100}{4 \cdot 74 - 16^2} = \frac{1800 - 1600}{296 - 256} = \frac{200}{40} = 5$$

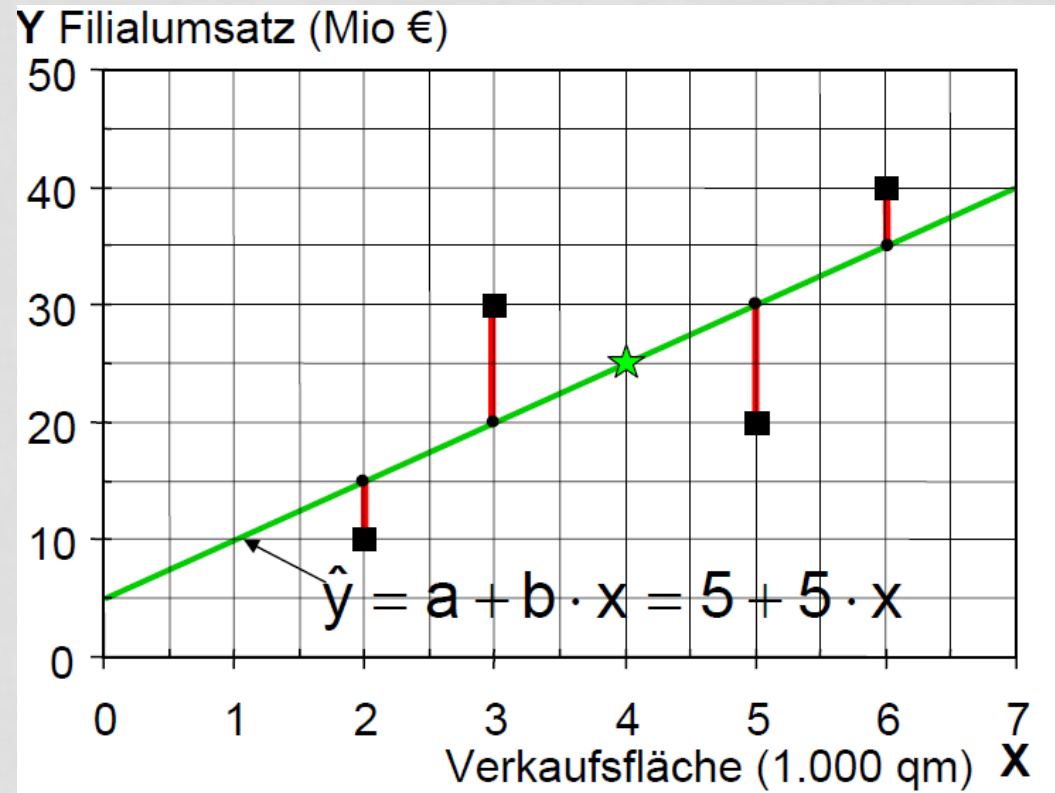
INTERPRETATION DER ERGEBNISSE

Regressionskoeffizienten a und b

Beispiel:

Fragen:

- Umsatzprognose für neue Filiale mit 4.500 qm
- Lohnt sich eine Erweiterung um 1.000 qm?



INTERPRETATION DER ERGEBNISSE

Regressionskoeffizienten a und b

Beispiel:

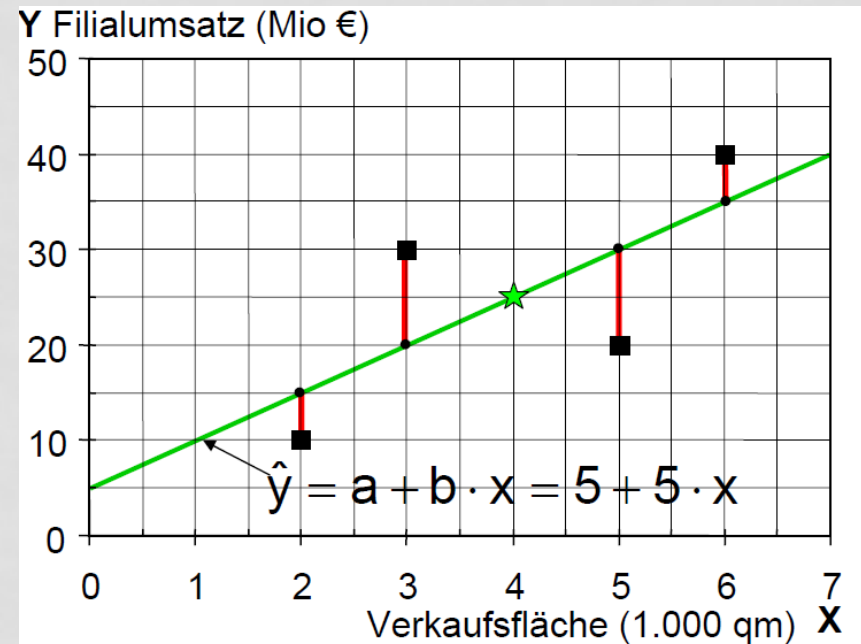
Fragen:

- Umsatzprognose für neue Filiale mit 4.500 qm Verkaufsfläche:

$$5 + 5 \cdot 4,5 = 27,5 \text{ Mio €}$$

- Lohnt sich eine Erweiterung um 1.000 qm Verkaufsfläche?

$$\text{mit Erweiterung: } 5 + 5 \cdot (4,5 + 1,0) = 5 + 5 \cdot 5,5 = 32,5 \text{ Mio €}$$

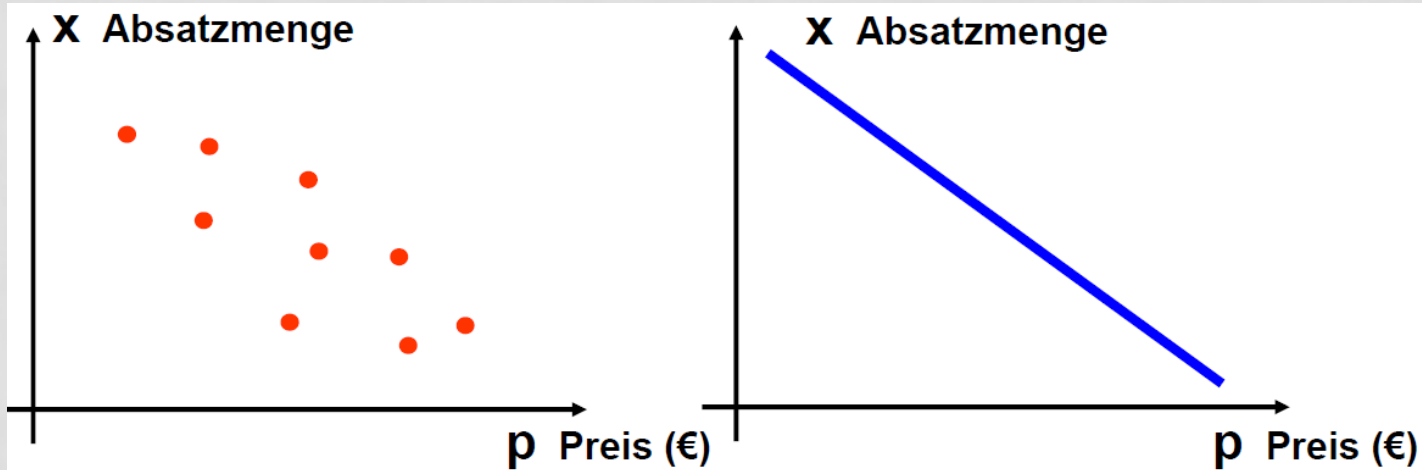


ANWENDUNG DER REGRESSIONSANALYSE

Regressionsverfahren haben viele praktische Anwendungen. Die meisten Anwendungen fallen in eine der folgenden beiden Kategorien:

- zum Erstellen eines **Vorhersagemodells**
- um die **Stärke des Zusammenhangs** zu quantifizieren: so können diejenigen x_j ermittelt werden, die gar keinen Zusammenhang mit y haben oder diejenigen Teilmengen x_i, \dots, x_j , die redundante Information über y enthalten.

MODELL VS. REALITÄT



Modell = vereinfachtes Abbild der Realität

- Wie gut beschreibt das Modell die Realität?
- Wie gut wird die Realität durch das Modell wiedergegeben?

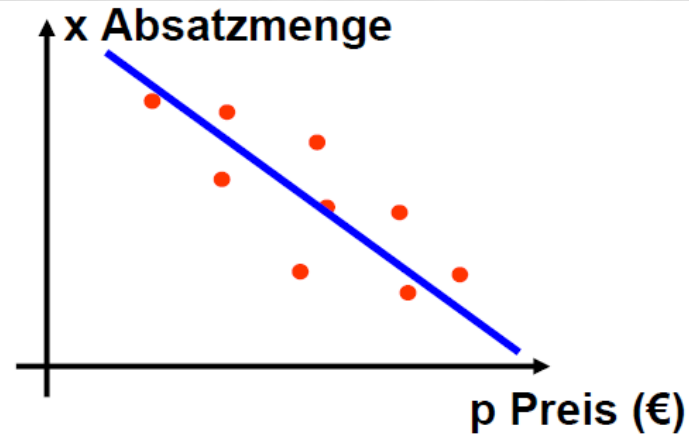
MODELL VS. REALITÄT

Regressionsrechnung:

$$\hat{x}(p) = a + b \cdot p = 100 - 5 \cdot p$$

Spezifikation
des Modells

Schätzung der
Parameter a, b



→ Wir brauchen **Gütemaße** für die Schätzung der Parameter:

- Wie gut ist die „goodness of fit“ (Anpassungsgüte)
- Wie gut beschreibt die Regressionsfunktion die Abhängigkeit?

MODELL VS. REALITÄT

Beispiel:

Verkaufsflächen

unterschiedlich groß



Filialumsatz

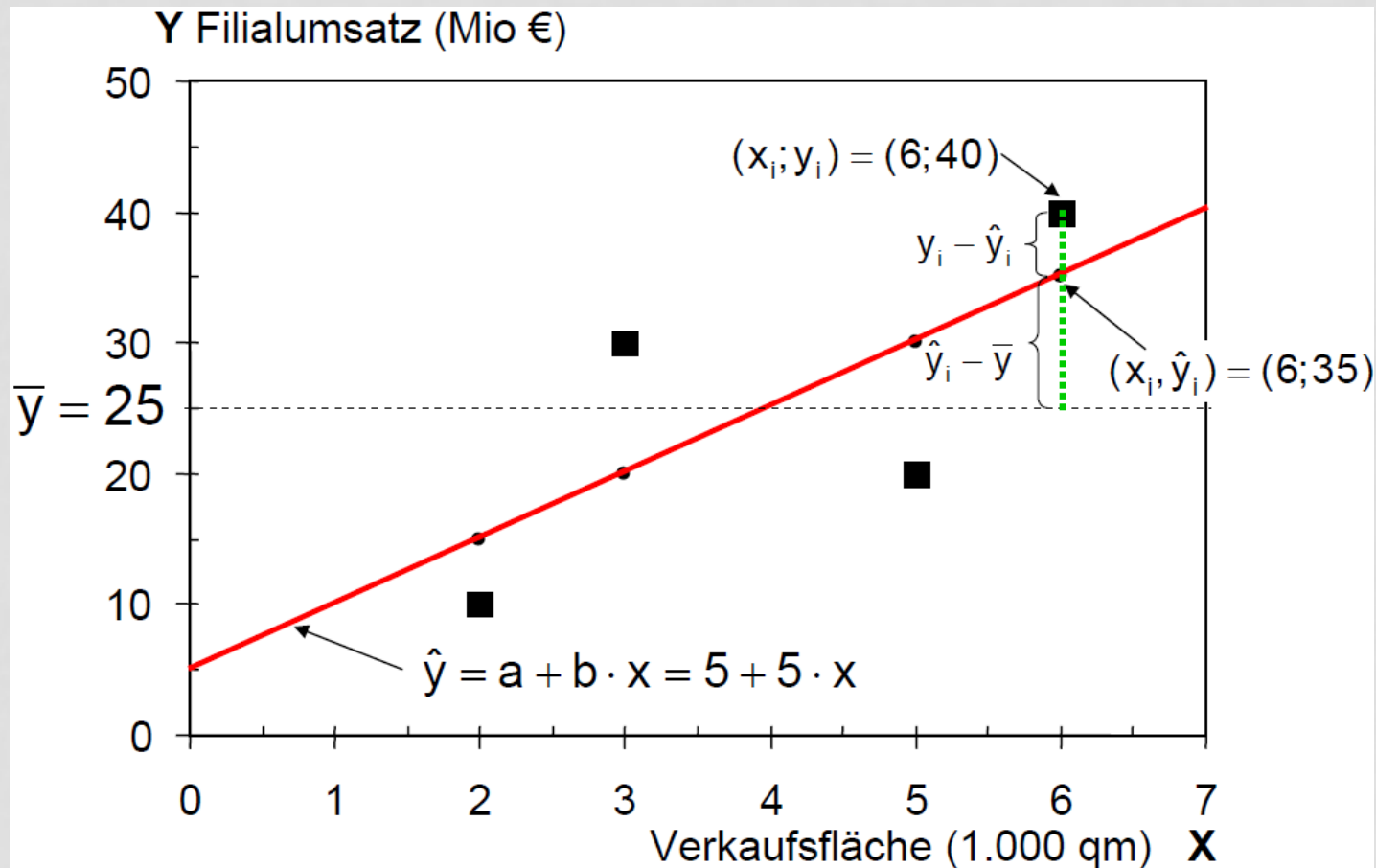
unterschiedlich hoch

WARUM?

- Wie gut erklären die Unterschiede bei den Verkaufsflächen die Unterschiede bei den Filialumsätzen?
 - **Wie viel Varianz wird durch das Modell nicht erklärt?**
- Wie gut erklärt die Regressionsfunktion die Abhängigkeit zwischen Verkaufsfläche und Filialumsatz?
 - **Wie hoch ist die Erklärungskraft des Modells?**

PROGNOSEWERTE UND RESIDUEN

Abweichungszerlegung: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$



(Vgl. mit Folie 17)

PROGNOSEWERTE UND RESIDUEN

Abweichungszerlegung: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Gesamtabweichung der Beobachtung y_i zum Mittelwert \bar{y}
→ diesen Fehler würden wir machen, wenn wir mit dem Mittelwert die entsprechende Beobachtung vorhersagen würden

Unerklärte Abweichung /Residuum der Beobachtung y_i zur Regressionsgeraden (Residuum u_i)
→ diesen Teil der Abweichung können wir auch durch Hinzunahme der unabhängigen Variablen x nicht vermeiden

Erklärte Abweichung der Regressionsgeraden zum Mittelwert \bar{y}
→ diesen Fehler können wir durch Hinzunahme der unabhängigen Variablen x vermeiden

Filiale Nr. 3:

$$\begin{array}{ccccccc} (40 - 25) & = & (40 - 35) & + & (35 - 25) & & \text{(Mio €)} \\ 15 & & = 5 & & + 10 & & \text{(Mio €)} \end{array}$$

PROGNOSEWERTE UND RESIDUEN

Varianzzerlegung:

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

$$s_y^2 = s_u^2 + s_{\hat{y}}^2$$

Die Varianz der Regressionswerte wird auch bestimmt durch die Varianz des unabhängigen Merkmals:

$$s_{\hat{y}}^2 = b^2 * s_x^2$$

PROGNOSEWERTE UND RESIDUEN

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2	\hat{y}_i
1	3	30	90	9	900	20
2	2	10	20	4	100	15
3	6	40	240	36	1.600	35
4	5	20	100	25	400	30
Summe	16	100	450	74	3.000	100

$$\hat{y} = 5 + 5 \cdot x$$

$$s_Y^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{1}{4} \cdot 3.000 - 25^2 = 750 - 625 = 125$$

$$s_X^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{1}{4} \cdot 74 - 4^2 = 18,5 - 16 = 2,5$$

$$s_{\hat{Y}}^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n \hat{y}_i^2 \right) - \bar{\hat{y}}^2 = \frac{1}{4} \cdot (20^2 + 15^2 + 35^2 + 30^2) - 25^2 = 687,5 - 625 = 62,5$$

PROGNOSEWERTE UND RESIDUEN

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2	\hat{y}_i	$u_i =$ $(y_i - \hat{y}_i)$
1	3	30	90	9	900	20	+10
2	2	10	20	4	100	15	-5
3	6	40	240	36	1.600	35	+5
4	5	20	100	25	400	30	-10
Summe	16	100	450	74	3.000	100	0

Varianz der Residuen:

$$s_u^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n u_i^2 \right) - \bar{u}^2 = \frac{1}{4} \cdot (10^2 + (-5)^2 + 5^2 + (-10)^2) - 0^2 =$$

$$\frac{1}{4} \cdot 250 - 0 = 62,5$$

BESTIMMTHEITSMAB

Das **BestimmtheitsmaB** R^2 (Erklärungskraft des Modells) ist ein **GütemaB der linearen Regression**.

Das R^2 gibt an, wie gut die unabhängige Variable Y geeignet ist, die Varianz der abhängigen Variable X zu erklären.

(unbrauchbares Modell) **0% $\leq R^2 \leq$ 100%** (perfekte Modellanpassung)

Das R^2 nutzt das Konzept der Varianzzerlegung und besagt, dass sich die Varianz des abhängigen Merkmals in erklärte Varianz und nicht erklärte Varianz (Residualvarianz) zerlegen lässt.

BestimmtheitsmaB R^2 \rightarrow Anteil der Varianz der abhängigen Variable, der sich durch die Varianz der unabhängigen Variable erklären lässt.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{erklärte Variation}}{\text{Gesamtvariation}} = 1 - \frac{\text{unerklärte Variation}}{\text{Gesamtvariation}} = \frac{\sum_{i=1}^n u_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

BESTIMMTHEITSMAB

Es folgt: R^2 ist das Verhältnis aus der Streuung der Prognosewerte und der Gesamtstreuung der y-Werte (**s. Folie 38**):

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{62,5}{125} = 0,5 \Leftrightarrow 1 - \frac{s_u^2}{s_y^2} = 1 - \frac{62,5}{125} = 1 - 0,5 = 0,5$$

Achtung!

→ Bei einer einfachen linearen Regression (nur eine unabhängige Variable) entspricht das Bestimmtheitsmaß dem Quadrat des Korrelationskoeffizienten nach Pearson r_{XY}

$$R^2 = (r_{XY})^2$$

Beispiel:

X = Verkaufsfläche, Y = Filialumsatz

$$r_{XY} = 0,707 \rightarrow R^2 = 0,707^2 = 0,50 = 50 \%$$

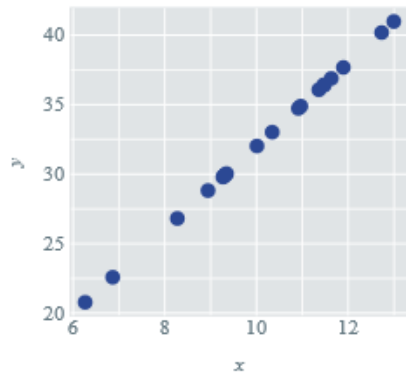
Bedeutung / Interpretation:

50 % der Varianz der Filialumsätze lassen sich durch die Varianz der Verkaufsflächen erklären. Die anderen 50 % lassen sich nur durch andere Einflussfaktoren erklären.

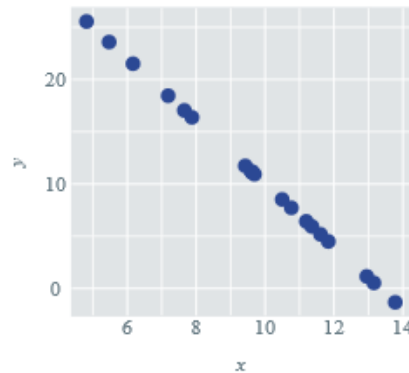
BESTIMMTHEITSMAB

Die folgende Grafiksammlung zeigt verschiedene Streudiagramme in Abhängigkeit des Wertes des R^2 . Je eher die Datenpunkte auf einer Linie liegen, desto höher ist das R^2 . Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt das R^2 nahe 0.

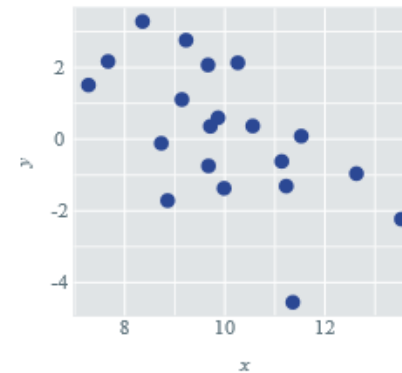
R = 1



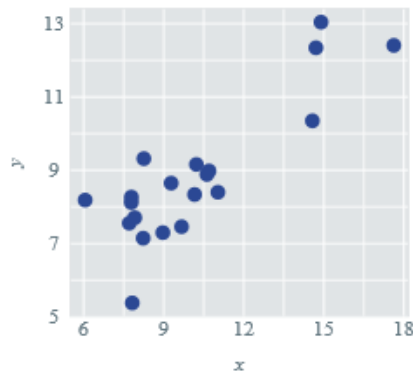
R = 1



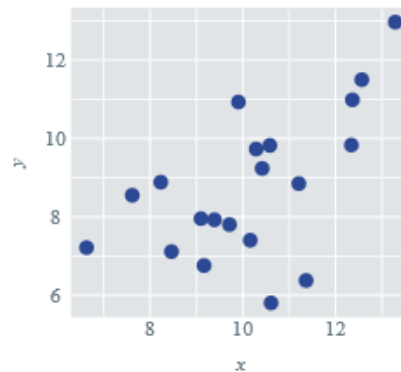
R = 0.36



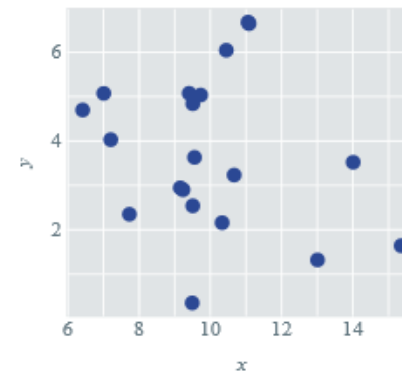
R = 0.73



R = 0.34



R = 0.05



BESTIMMTHEITSMAB

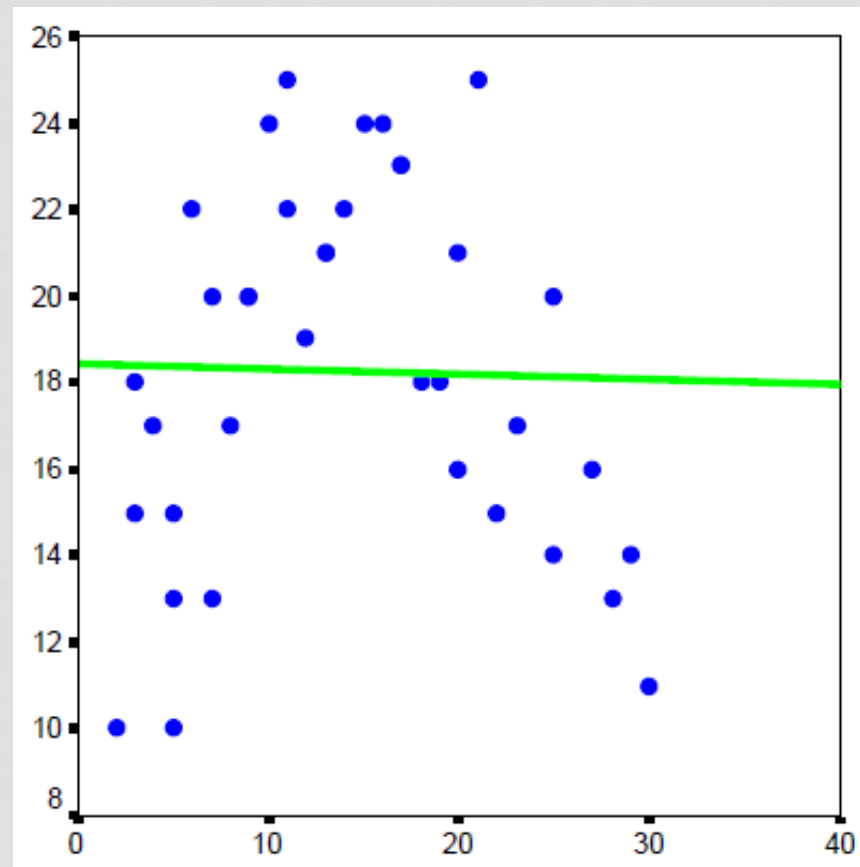
"Wie hoch muss mein R^2 sein?"

- Die übliche Größenordnung des R^2 variiert, je nach dem um welches Anwendungsgebiet es sich handelt. In Bereichen wie dem klassischen Marketing, in denen es hauptsächlich darum geht, menschliches Verhalten zu erklären bzw. vorherzusagen, sind meist geringe R^2 (deutlich kleiner 50%) zu erwarten. In anderen Bereichen wie bspw. der Physik sind höhere R^2 die Regel. Dies ist wenig überraschend, da auf das menschliche Verhalten zahlreiche und häufig nicht direkt messbare Einflüsse wirken. In der Physik hingegen werden oft Zusammenhänge zwischen wenigen exakt messbaren Größen untersucht. Dies geschieht zusätzlich meist unter experimentellen Bedingungen, unter denen sich Störeinflüsse minimieren lassen.
- Während auf der Mikro-Ebene in vielen Fällen bereits ein R^2 von 10% als gut gelten kann, erwarten viele bei stärker aggregierten Daten ein R^2 von 40% bis 80% oder sogar mehr. Ein Modell mit geringem R^2 - selbst bei stärker aggregierten Daten – ist nicht nutzlos, da die Alternative dazu oft gar kein Modell darstellt, was einem R^2 von 0 entspricht. Im übertragenen Sinne bedeutet das, dass eine systematische Prognose auf Basis eines Modells mit beschränktem R^2 oft schon besser ist als eine unsystematische Planung, die ausschließlich auf Bauchgefühl setzt. Generell ist die Aussagekraft von Modellen mit geringem R^2 nicht zwangsläufig schlecht.

PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

Nur lineare Zusammenhänge werden erfasst!

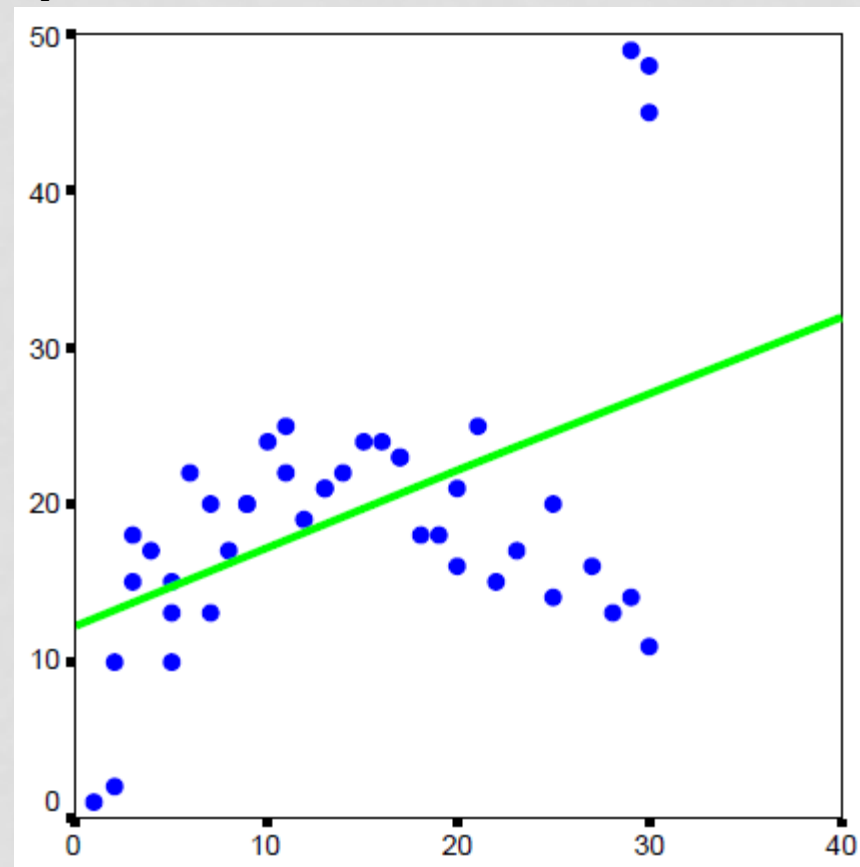
Die Gerade ist quasi horizontal – was nicht dem „eigentlichen“ Zusammenhang entspricht. Hier geht es um einen nicht linearen Zusammenhang, der nicht durch die lineare Regression beschrieben werden kann.



PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

**Einzelne Fälle können starken Einfluss ausüben
(nicht zuletzt wegen dem Quadrieren)!**

Die gleichen Daten wie vorhin
plus einige Extremwerte
(links unten, rechts oben)
erzeugen eine deutlich
steigende Gerade



PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

**Einzelne Fälle können starken Einfluss ausüben
(nicht zuletzt wegen dem Multiplizieren)!**

Korrelation über alle Fälle: $r=0,35$

Korrelation ohne Extremfall

(y über 14.000): $r=0,39$

