Machine Learning

Prof. Dr. Fabian Brunner

<fa.brunner@oth-aw.de>

Amberg, 7. November 2023

Übersicht

Ostbayerische Technische Hochschule
Amberg-Weiden

Themen letzte Woche:

- Lineare Regression
- Mathematische Grundlagen
- Least Squares Funktional
- Normalgleichungen
- Gradientenverfahren

Themen heute:

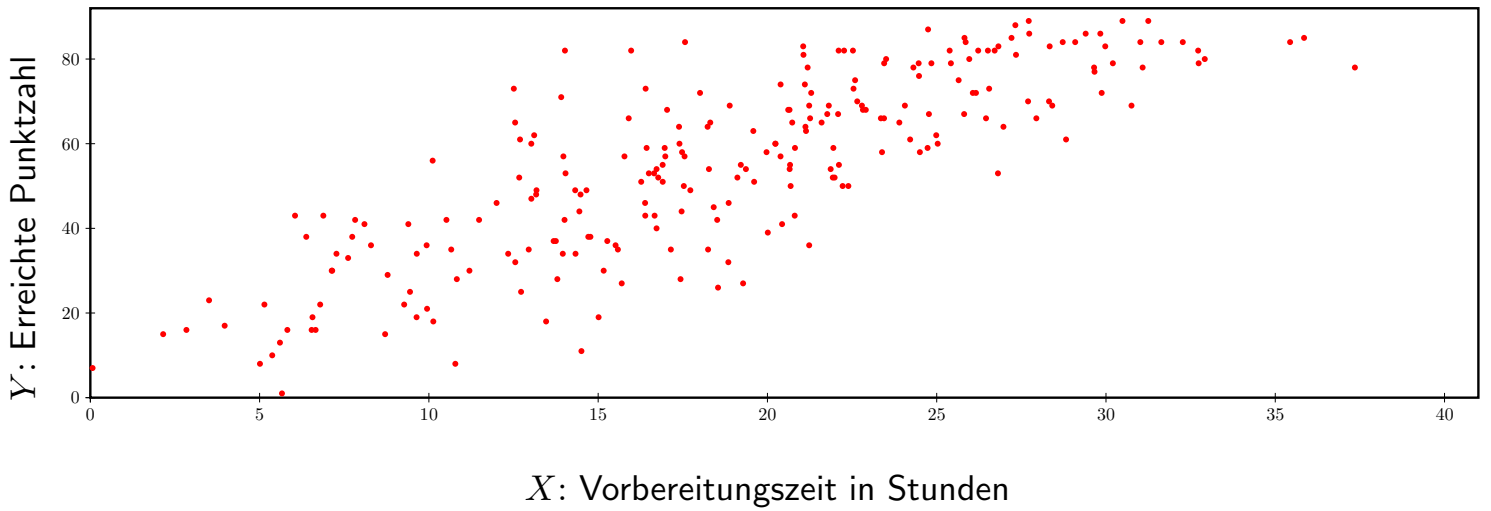
- Korrelationsanalyse
- RANSAC-Regression
- Bewertung der Güte von linearen Regressionsmodellen
- Signifikanz von linearen Regressionsmodellen

Korrelationsanalyse

Häufig betrachtet man für ein Zufallsexperiment mehrere Zufallsvariablen oder Zufallsvariablen, die aus mehreren Zufallsvariablen gebildet werden können:

- Körpergröße und Gewicht einer zufällig aus einer Population gezogenen Person
- Alter und Ausfallhäufigkeit von Maschinen
- Vorbereitungszeit und Prüfungsergebnis eines zufällig ausgewählten Studenten.

Korrelationsanalyse: Untersuchung und Quantifizierung von linearen Zusammenhängen zwischen Zufallsgrößen



Die Kovarianz

Definition der Kovarianz

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$. Dann ist ihre Kovarianz gegeben durch

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) \ .$$

Die Kovarianz

Definition der Kovarianz

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$. Dann ist ihre Kovarianz gegeben durch

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) \ .$$

Beispiel:
Zwei Zufallsvariablen X und Y haben die folgende gemeinsame Wahrscheinlichkeitsfunktion:

		Y			
		1	2	3	$f_X(x)$
X	1	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
	2	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
	$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

mit $\mu_X = 1.5$, $\mu_Y = 2$, $\text{Var}(X) = \frac{1}{4}$, $\text{Var}(Y) = \frac{1}{2}$. Dann lautet die Kovarianz

Die Kovarianz

Definition der Kovarianz

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$. Dann ist ihre Kovarianz gegeben durch

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) \ .$$

Beispiel:
Zwei Zufallsvariablen X und Y haben die folgende gemeinsame Wahrscheinlichkeitsfunktion:

		Y			
		1	2	3	$f_X(x)$
X	1	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
	2	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
	$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

mit $\mu_X = 1.5$, $\mu_Y = 2$, $\text{Var}(X) = \frac{1}{4}$, $\text{Var}(Y) = \frac{1}{2}$. Dann lautet die Kovarianz

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = \sum_{i=1}^{2} \sum_{j=1}^{3} (x_i - \frac{3}{2})(y_j - 2) f(x_i, y_j) = \ldots = \frac{1}{4} \ .$$

Eigenschaften der Kovarianz

Eigenschaften der Kovarianz

1. $\mathrm{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$ (alternative Berechnungsformel)
2. X, Y stochastisch unabhängig $\Rightarrow \mathrm{Cov}(X, Y) = 0$

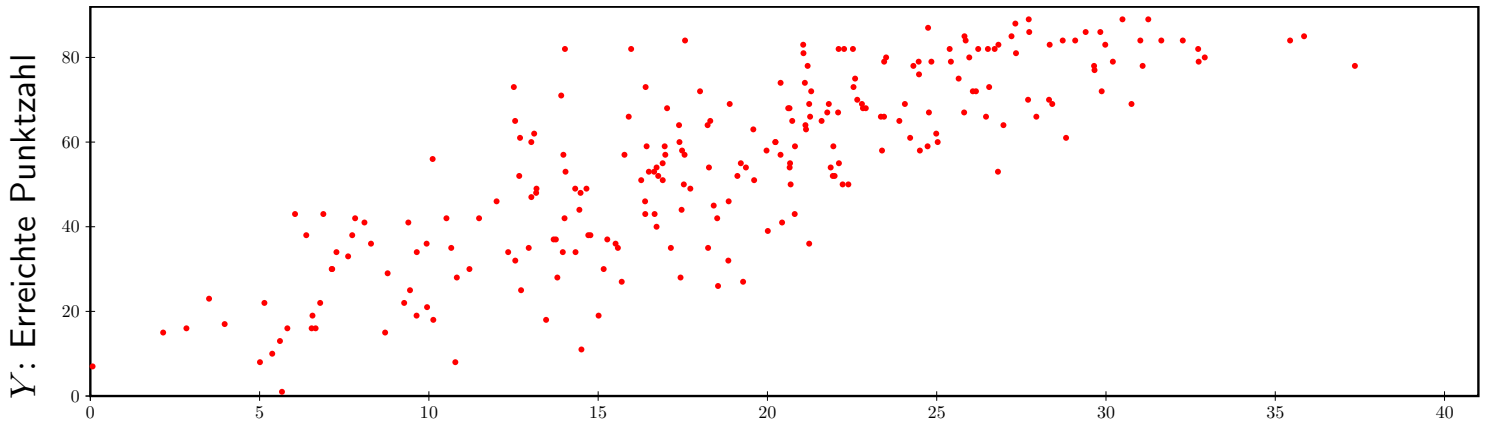
   Cov misst nur lineare Unabhängigkeit
3. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
4. $|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}$ (vgl. Korrelationskoeffizient)
5. Symmetrie und Linearität:
   - $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
   - $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)) = \mathrm{Cov}(X, aY)$ für alle $a \in \mathbb{R}$
   - $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$

Bemerkung: Die Umkehrung von 2. gilt im Allgemeinen nicht!

Das Vorzeichen der Kovarianz

Zurück zum Eingangsbeispiel: die Stichproben (x_i, y_i) stammen aus Verteilungen mit $\mu_X = 20$, $\mu_Y = 60$, und

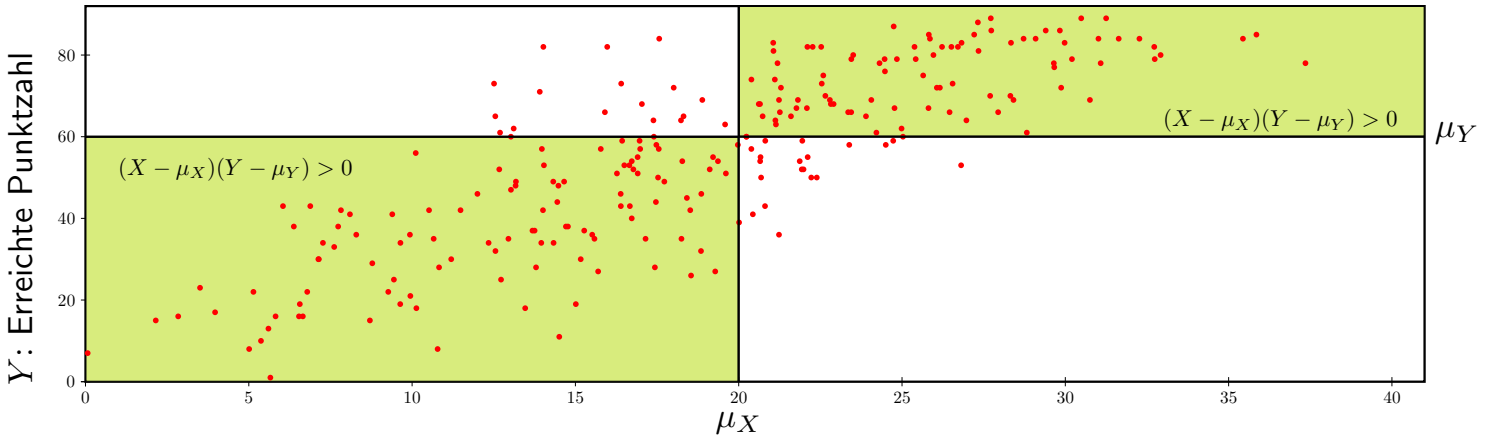
$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = +240 \ .$$



Das Vorzeichen der Kovarianz

Zurück zum Eingangsbeispiel: die Stichproben (x_i, y_i) stammen aus Verteilungen mit $\mu_X = 20$, $\mu_Y = 60$, und

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = +240 \ .$$

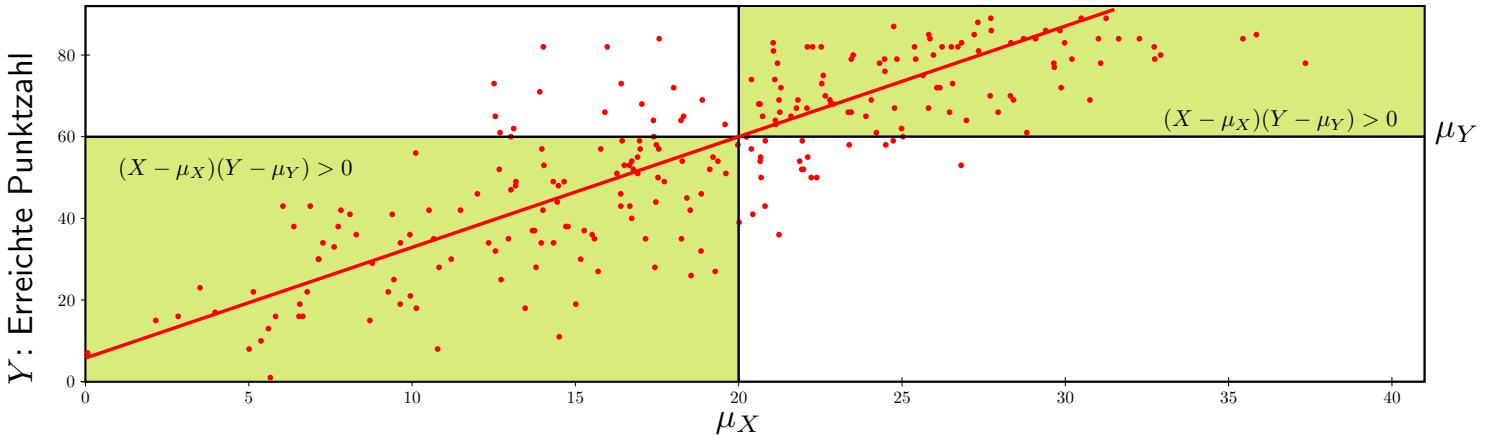


mehrheit der Punkte in grünen kästen
steigender zusammenhang → große X → große Y
→ einheiten loswerden durch $VAR(X)$ und $VAR(Y)$ teilen siehe F.5

Das Vorzeichen der Kovarianz

Zurück zum Eingangsbeispiel: die Stichproben (x_i, y_i) stammen aus Verteilungen mit $\mu_X = 20$, $\mu_Y = 60$, und

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = +240 \ .$$



X: Vorbereitungszeit in Stunden

Das Vorzeichen der Kovarianz

- Ist $\mathrm{Cov}(X, Y) > 0$, so treten hohe Werte von X tendenziell mit hohen Werten von Y auf.
- Ist $\mathrm{Cov}(X, Y) < 0$, so treten hohe Werte von X tendenziell mit niedrigen Werten von Y auf.

Der Korrelationskoeffizient

Definition 1

Seien X und Y zwei gemeinsam verteilte Zufallsvariablen mit Standardabweichungen $\sigma_X = \sqrt{\mathrm{Var}(X)} > 0$ und $\sigma_Y = \sqrt{\mathrm{Var}(Y)} > 0$. Dann ist der Korrelationskoeffizient von X und Y definiert durch

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \ .$$

Eigenschaften des Korrelationskoeffizienten

Der Korrelationskoeffizient ist ein Maß für die Art und Stärke eines linearen Zusammenhangs zwischen zwei Zufallsvariablen X und Y:

Eigenschaften des Korrelationskoeffizienten

1. Normiertheit: $-1 \leq \rho_{X,Y} \leq 1$.
2. Perfekter linearer Zusammenhang:
   - Ist $Y = aX + b$ mit Konstanten $a, b \in \mathbb{R}$, $a \neq 0$, dann gilt

$$\rho_{X,Y} = \begin{cases} +1 & \text{falls } a > 0 \ , \\ -1 & \text{falls } a < 0 \ . \end{cases}$$

   - Ist $|\rho_{X,Y}| = 1$, dann gibt es Konstanten a, $b \in \mathbb{R}$, sodass

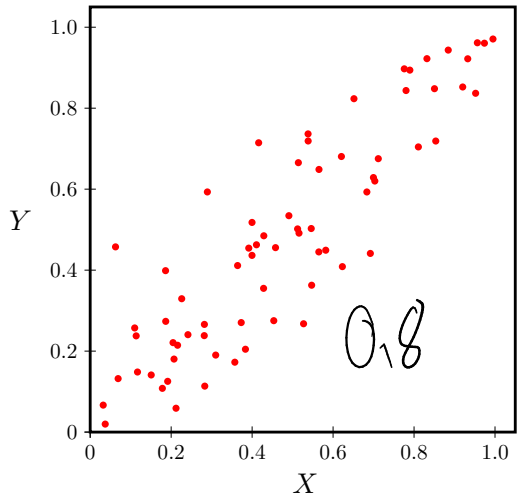
$$P(Y = aX + b) = 1 \ .$$

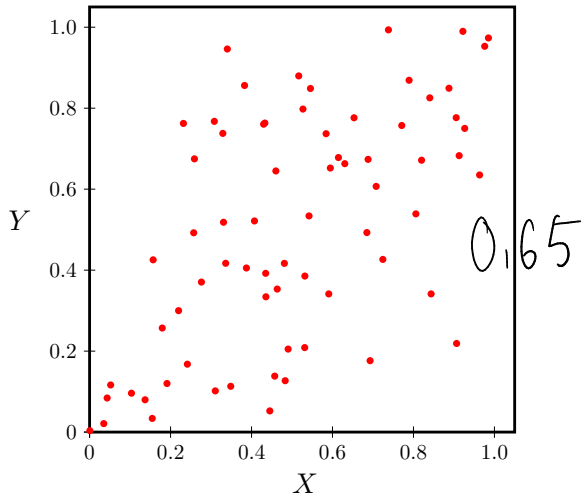
   Für $\rho_{X,Y} = +1$ gilt $a > 0$ und für $\rho_{X,Y} = -1$ gilt $a < 0$.

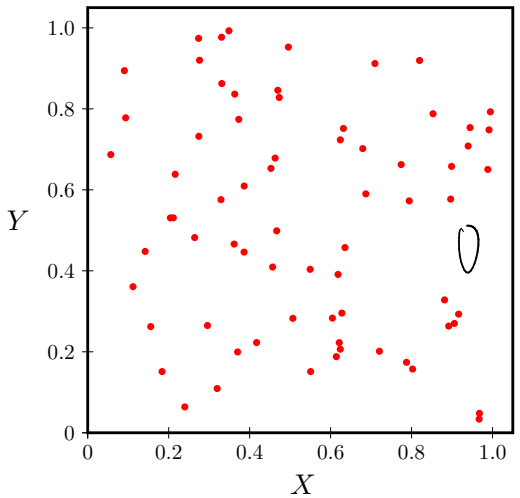
Bemerkung: Je näher $|\rho_{X,Y}|$ an 1 liegt, desto stärker ausgeprägt ist der lineare Zusammenhang zwischen X und Y.

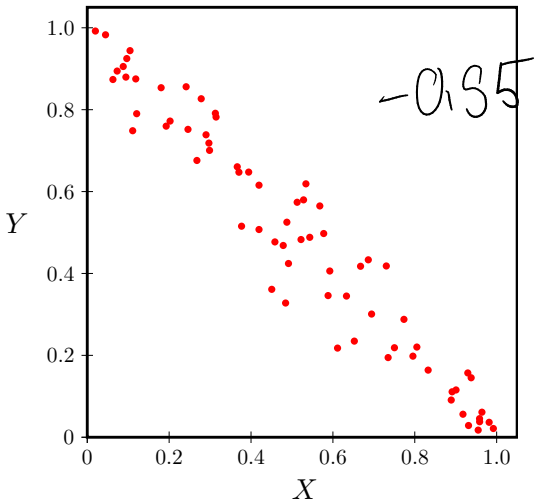
Interpretation des Korrelationskoeffizienten

Die nachfolgenden Stichproben stammen aus zweidimensionalen Verteilungen mit den Korrelationskoeffizienten $\rho_{X,Y} \in \{-0.95,\ 0,\ 0.65,\ 0.8\}$. Wie lautet die korrekte Zuordnung?



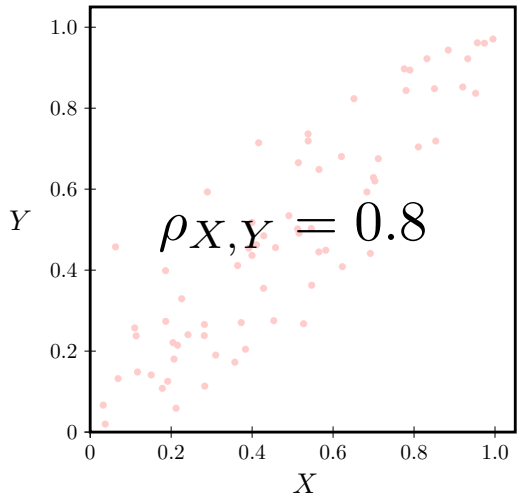


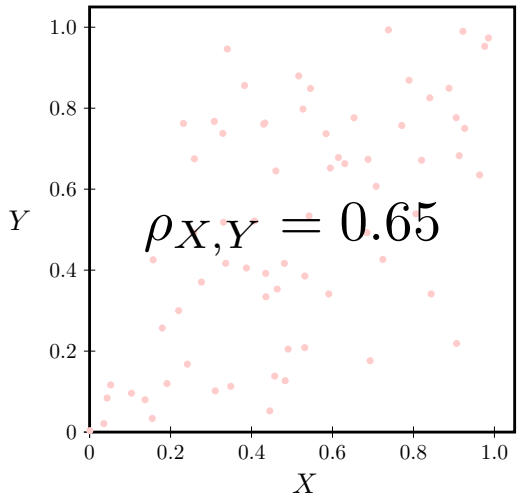


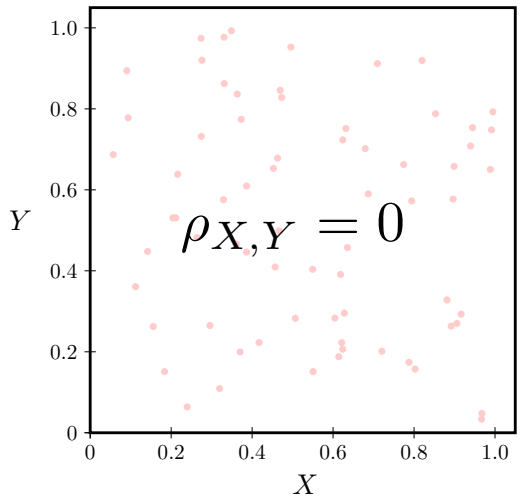


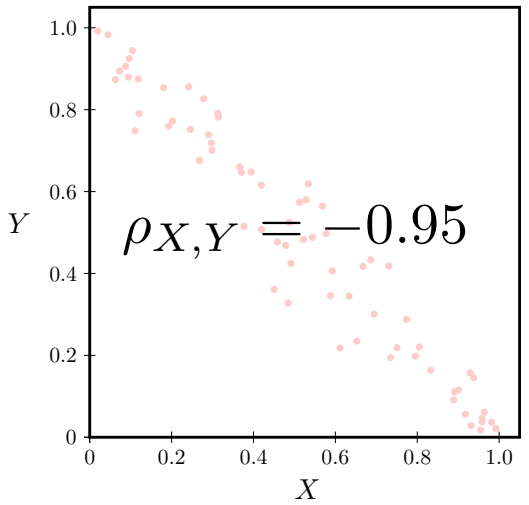
Interpretation des Korrelationskoeffizienten

Die nachfolgenden Stichproben stammen aus zweidimensionalen Verteilungen mit den Korrelationskoeffizienten $\rho_{X,Y} \in \{-0.95,\ 0,\ 0.65,\ 0.8\}$. Wie lautet die korrekte Zuordnung?









Korrelation und Kausalität

Beachte:

- Ein hoher Korrelationskoeffizient zwischen zwei Zufallsvariablen impliziert nicht, dass zwischen ihnen ein kausaler Zusammenhang besteht.
- Häufig gibt es jedoch eine dritte Variable, die beide beeinflusst.

Beispiel: Korrelation zwischen Schokoladenkonsum und Anzahl der Nobelpreisträger pro Einwohner verschiedener Länder:

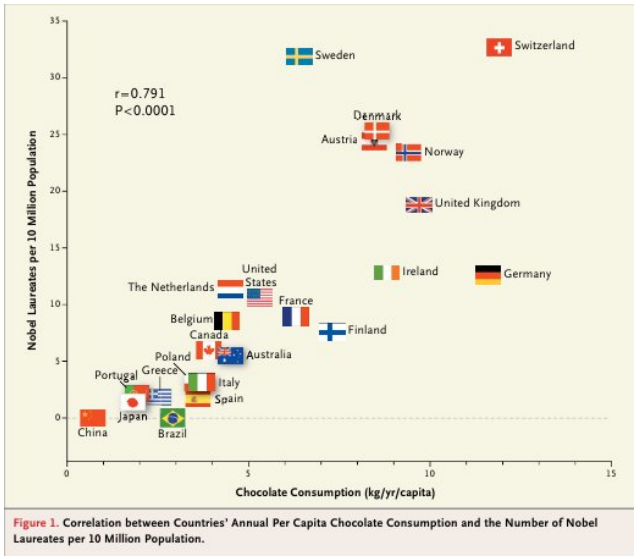


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

F. H. Messerli; New Engl. J. Med. 1562-1564; 2012

Korrelationsmatrix

Sind $X_1, X_2, X_3, \ldots, X_n$ Zufallsvariablen über einem gemeinsamen Grundraum Ω, dann enthält die **Korrelationsmatrix** die paarweise gebildeten Korrelationskoeffizienten:

$$R = \begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & \rho_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & 1 \end{pmatrix} .$$

Auf der Diagonalen stehen Einsen, da dort die Korrelationskoeffizienten der Zufallsvariablen mit sich selbst stehen und jede Zufallsvariable perfekt mit sich selbst korreliert ist.

Empirischer Korrelationskoeffizient

- Bei der Analyse von Daten ist die theoretische Verteilung der einzelnen Merkmale oftmals unbekannt.
- Stattdessen ist eine Stichprobe von Werten gegeben.
- Auf den Stichproben kann man einen empirischen Korrelationskoeffizienten bestimmen, indem man die Kovarianzen und Varianzen in der Definition des Korrelationskoeffizienten durch die entsprechenden empirischen Größen ersetzt.

Empirischer Korrelationskoeffizient nach Pearson

Für zwei Stichproben $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(m)})$ und $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(m)})$ zweier Merkmale ist der empirische Korrelationskoeffizient nach Pearson definiert durch

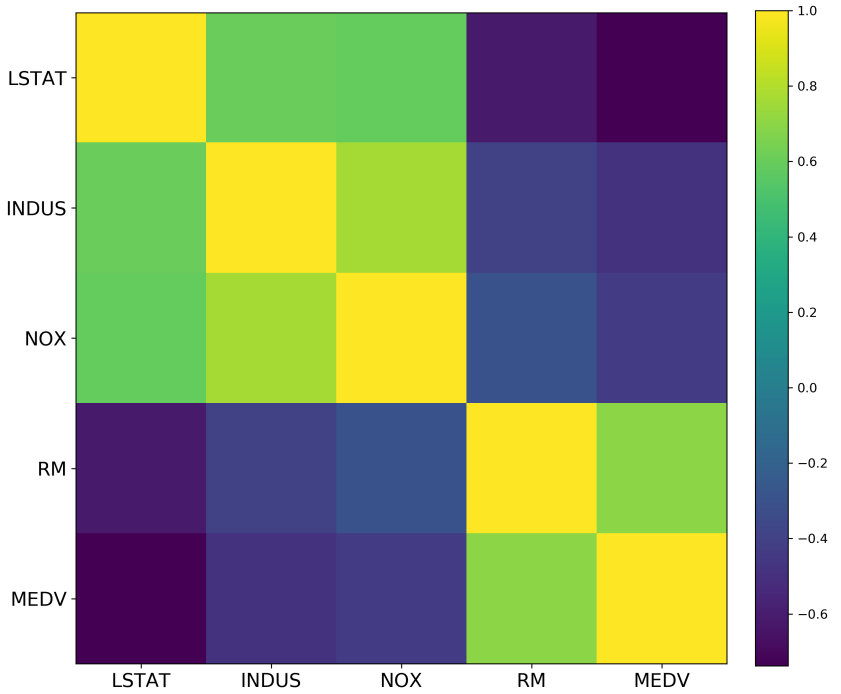
$$r_{\boldsymbol{xy}} := \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2 \cdot \sum_{i=1}^{m}(y^{(i)} - \bar{y})^2}} \,,$$

kürzt sich

wobei $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ und $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}$ die emprischen Stichprobenmittelwerte von $\boldsymbol{x}$ und $\boldsymbol{y}$ bezeichnen.

Correlation Heat Map

Die (empirische) Korrelationsmatrix wird häufig in Form einer Heat Map visualisiert. In der folgenden Korrelationsmatrix werden die empirischen Korrelationskoeffizienten der Spalten LSTAT, INDUS, NOX, RM und MEDV des „Boston Housing“- Datensatzes (vgl. Übung) dargestellt:

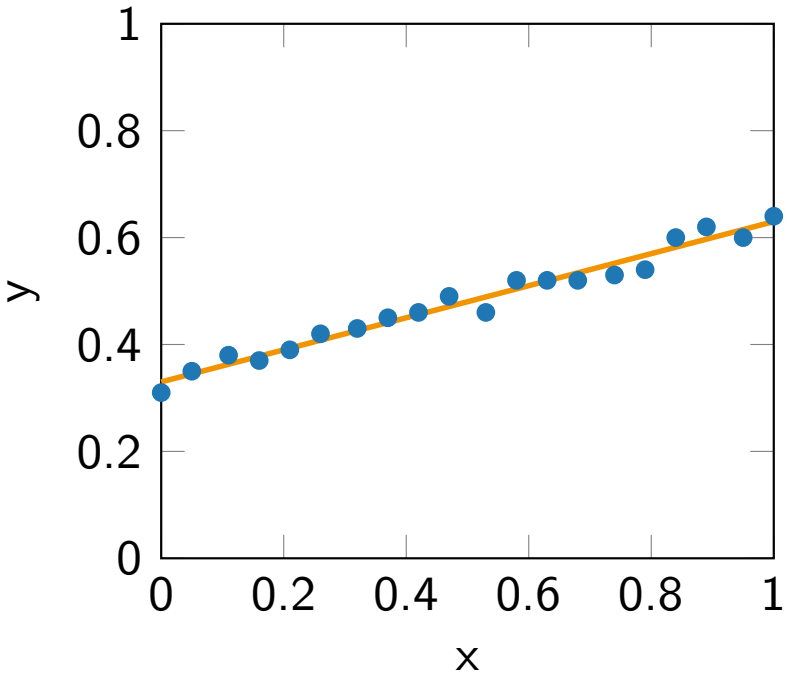


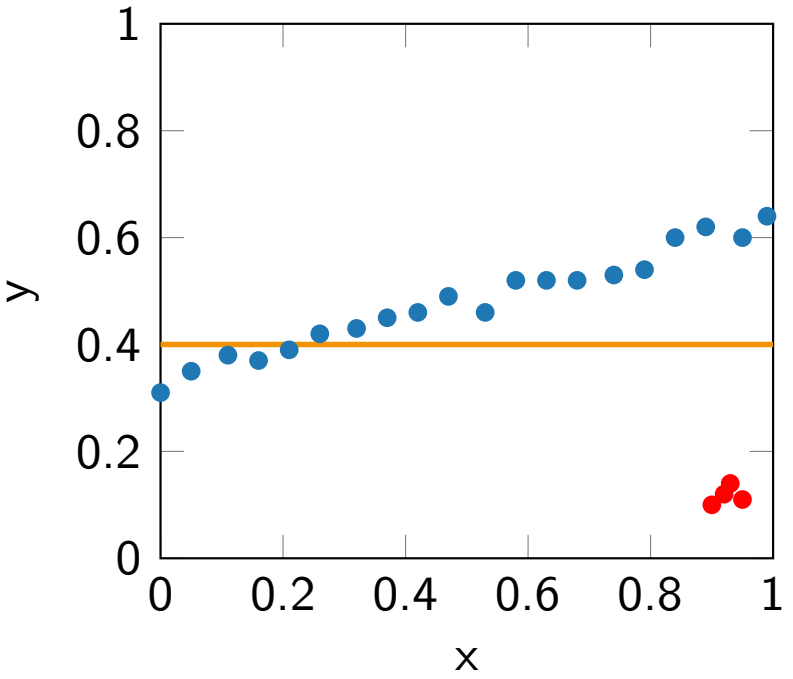
Zusammenfassung Korrelationsanalyse

- Der Korrelationskoeffizient misst die Art und Stärke eines linearen Zusammenhangs zweier Zufallsgrößen.
- Er gibt einen Anhaltspunkt, ob eine Modellierung mittels linearer Regression geeignet sein kann.
- Für eine Stichprobe von Daten kann der empirische Korrelationskoeffizient nach Pearson berechnet werden.
- Die paarweisen Korrelationskoeffizienten zwischen mehreren Zufallsvariablen bilden die Korrelationsmatrix.
- Diese wird häufig als Heat Map visualisiert.

Einfluss von Ausreißern

Lineare Regressionsmodelle sind nicht robust gegenüber Ausreißern („Outliers“) und können durch deren Vorhandensein stark verzerrt werden:





Ansätze

- Vorgelagerte Entfernung von Ausreißern
- Einsatz eines robusten Regressionsverfahrens → RANSAC

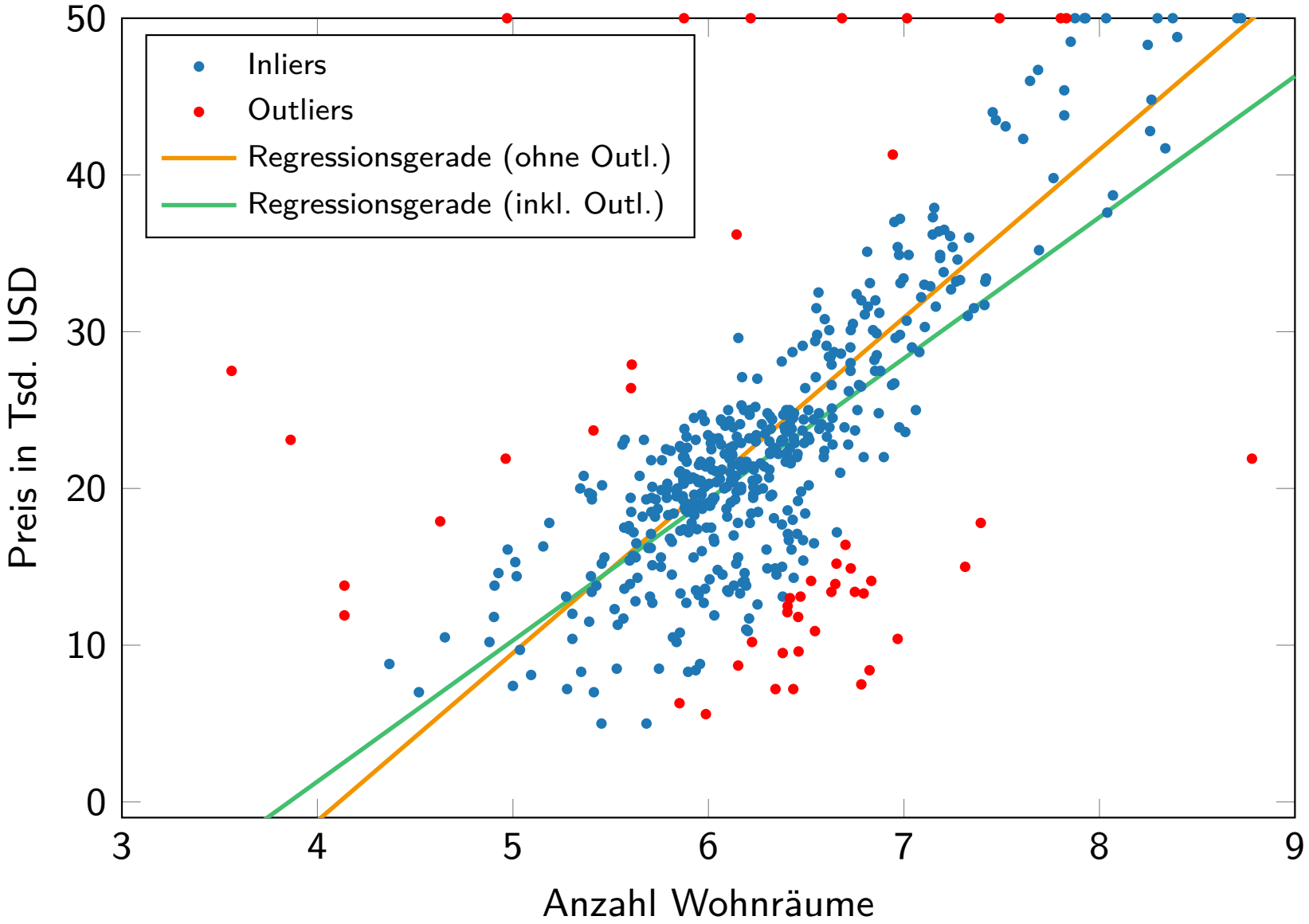
RANSAC-Algorithmus

RANSAC = „**RAN**dom **SA**mple **C**onsensus“

RANSAC

1. Auswahl einer zufälligen Stichprobe der Daten („Inliers“) und Modell-Fitting auf dieser Stichprobe.
2. Berechnung der Residuen für die anderen Datenpunkte
3. Hinzunahme derjenigen Punkte zu den Inliers, deren Residuum unterhalb eines definierten Schwellwerts liegt.
4. Modell-Fitting auf den in den Schritten 1-3 ermittelten Inliers („consensus set“).
5. Wiederhole die Schritte 1-3 solange bis ein Abbruchkriterium erfüllt ist (z.B. vorgegebene Anzahl an Iterationen, Mindestanzahl an Inliers).

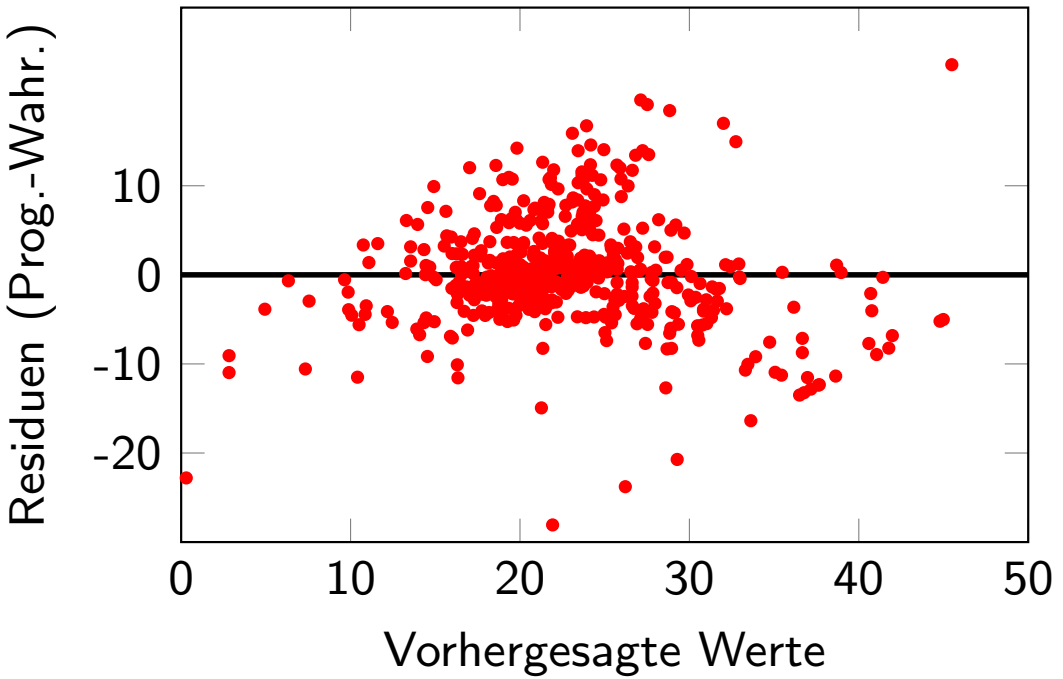
Beispiel zum RANSAC-Verfahren



Bewertung der Güte linearer Regressionsmodelle

Residuenberechnung:

- Eine Möglichkeit, die Güte linearer Regressionsmodelle zu bewerten, besteht darin, die Residuen für alle Punkte zu berechnen und darzustellen.
- Das Residuum ist die Abweichung zwischen dem wahren Wert und dem durch das Regressionsmodell prognostizierten Wert (vorzeichenbehaftet!).
- Zur grafischen Darstellung dienen Residualdiagramme (nützlich v.a. für multivariate Regression):



Bewertung der Leistung linearer Regressionsmodelle

- Die Beurteilung der Modellgüte anhand der Residuuen für alle Datenpunkte ist problematisch, da nicht gemessen werden kann, wie gut die Modellprognosen auf unbekannten Daten sind.
- Wir werden später noch sehen, dass viele Modelle auf den Trainingsdaten sehr gute Ergebnisse liefern, aber schlecht auf neue Daten verallgemeinern ($\rightarrow$ siehe Bias-Varianz-Zerlegung für Regressionsmodelle).
- Deshalb wird die Modellgüte typischerweise nicht auf dem Trainingsdatensatz ausgewertet, sondern auf einem unabhängigen Testdatensatz.

Merke:

Um einzuschätzen, wie gut ein Modell auf ungesehenen Daten funktioniert, sollte ein Machine Learning-Modell stets auf einer unabhängigen Testdatenmenge bewertet werden, deren Datensätze nicht beim Modelltraining verwendet wurden.

Mean squared error

- Der **Mean squared error** (kurz: MSE) ist ein quantitatives Maß zur Bewertung der Leistung eines Modells:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (f_{\theta}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$$

- Diese Kennzahl kann auf dem Trainings- und auf dem Testdatensatz ausgewertet und verglichen werden, um zu bewerten, wie gut das Modell auf unbekannten Daten verallgemeinert.
- Ist der MSE auf den Trainingsdaten deutlich geringer als auf den Testdaten, so deutet dies auf eine Überanpassung des Modells („Übertraining“) hin.

Mean absolut error: $MAE = \frac{1}{m} \sum_{i=1}^{m} (|f_{\theta}(x^{(i)}) - y^{(i)}|)^2$

Bestimmtheitsmaß

Eine weitere Größe zur Bewertung von Regressionsmodellen ist das Bestimmtheitsmaß R^2. Um dieses zu definieren, betrachten wir zunächst die folgenden Hilfsgrößen:

$$RSS = \sum_{i=1}^{m} (f_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)})^2 \ , \qquad \text{(Residual sum of squares)}$$

$$TSS = \sum_{i=1}^{m} (y^{(i)} - \bar{y})^2 \ , \qquad \text{(Total sum of squares)}$$

$$ESS = \sum_{i=1}^{m} (f_{\boldsymbol{\theta}}(x^{(i)}) - \bar{y})^2 \ , \qquad \text{(Explained sum of squares)}$$

wobei $\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y^{(i)}$ das arithmetische Mittel der Zielvariable auf den Trainingsdaten bezeichne.

Zerlegung von *TSS*

Theorem 2

Im Fall der Least-Squares-Regression gilt auf den Trainingsdaten die Beziehung

$$TSS = RSS + ESS \ .$$

Beweis: Sei $\hat{y}^{(i)} := f_{\boldsymbol{\theta}}(x^{(i)})$. Dann gilt

$$\begin{aligned} TSS &= \sum_{i=1}^{m} (y^{(i)} - \bar{y})^2 = \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)} + \hat{y}^{(i)} - \bar{y})^2 \\ &= \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^{m} (\hat{y}^{(i)} - \bar{y})^2 + 2 \underbrace{\sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y})}_{= \sum_{i=1}^{m} e^{(i)} (\hat{y}^{(i)} - \bar{y}) \overset{(1),(2),(3)}{=} 0} \\ &= RSS + ESS \ . \end{aligned}$$

Bestimmtheitsmaß (R^2)

Nun können wir das Bestimmtheitsmaß definieren, die als Maß für die Anpassungsgüte eines Regressionsmodells gilt:

Definition 3 (Bestimmtheitsmaß)

Das Bestimmtheitsmaß R^2 ist definiert durch

$$R^2 := \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \, .$$

Bemerkungen:

- Es gilt $R^2 = 1 - \frac{MSE}{\hat{\sigma}_y}$, wobei $\hat{\sigma}_y$ die empirische Standardabweichung der Zielvariablen bezeichnet.
- Auf den Trainingsdaten gilt $TSS = RSS + ESS$, d.h.

$$R^2 = \frac{ESS}{TSS} \, .$$

  Das Bestimmtheitsmaß gibt also an, welcher Anteil der Variation in den Daten durch das Modell erklärt wird.

Wertebereich von R^2

Ostbayerische Technische Hochschule Amberg-Weiden

Bemerkungen zum Wertebereich

- Aus der Definition von R^2 folgt unmittelbar die obere Schranke

$$R^2 \leq 1$$

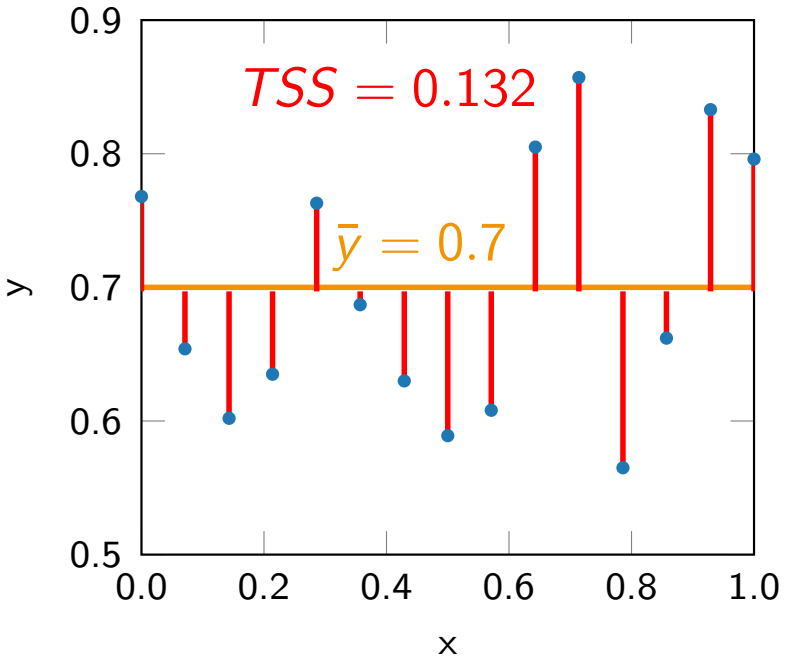
für den Wertebereich des Bestimmtheitsmaßes.

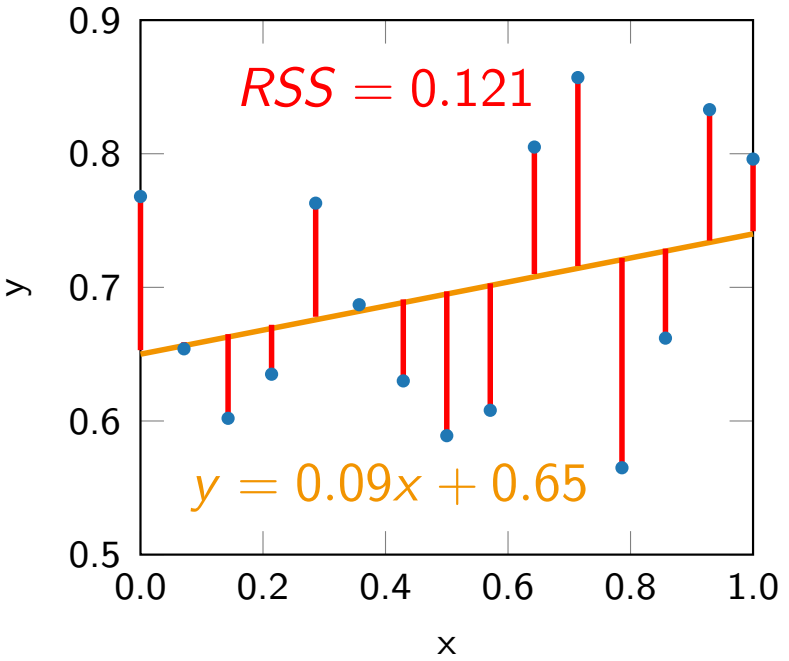
- Wegen $R^2 = \frac{ESS}{TSS}$ gilt auf dem Trainingsdatensatz außerdem

$$R^2 \geq 0 \ .$$

- Je näher der R^2 an 1 liegt, desto besser ist tendenziell die Anpassungsgüte des Regressionsmodells.
- Verständnisfrage: Wie sieht ein Modell aus, für das $R^2 = 0$ auf dem Trainingsdatensatz gilt?

Beispiel zum Bestimmtheitsmaß



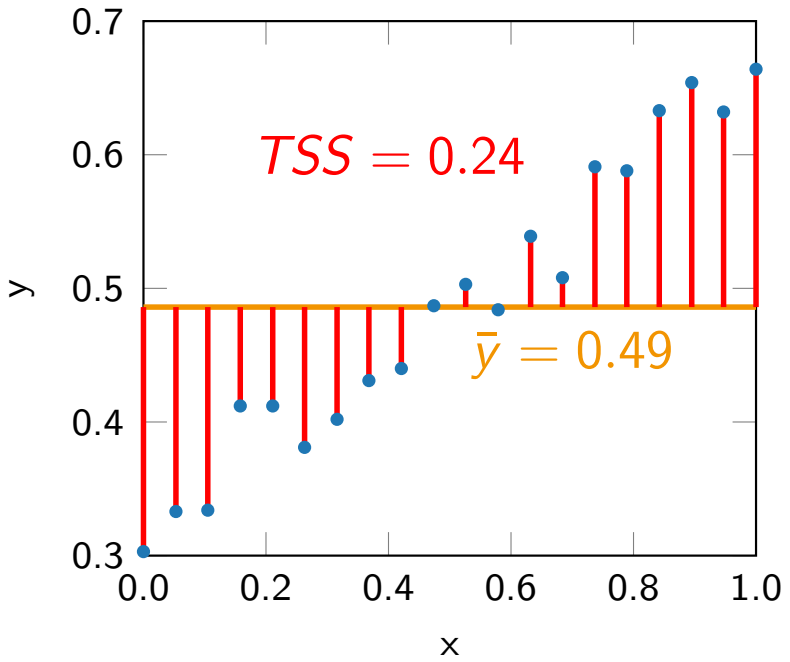


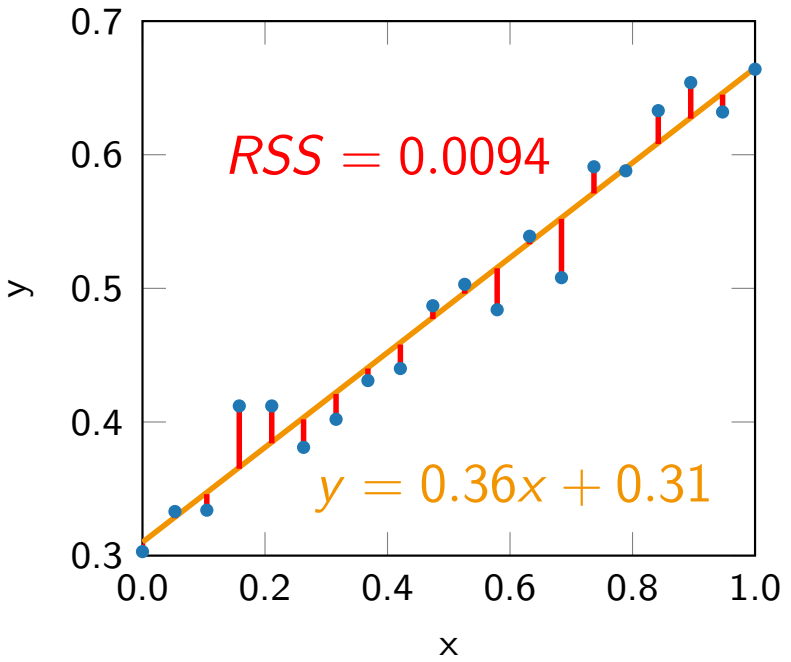
Man erhält

$$R^2 = \frac{TSS - RSS}{TSS} = 0.083 \ ,$$

d.h. die gefittete Gerade erklärt 8.3% der Varianz der Daten.

Beispiel zum Bestimmtheitsmaß





Man erhält

$$R^2 = \frac{TSS - RSS}{TSS} = 0.96 \ ,$$

d.h. die gefittete Gerade erklärt 96% der Varianz der Daten.

Bezug zum Pearson'schon Korrelationskoeffizienten

In Fällen, in denen (1), (2) und (3) aus Theorem 7 (vgl. letzte Vorlesung) gilt, entspricht das Bestimmtheitsmaß dem Quadrat des Pearson'schen Korrelationskoeffizienten zwischen $\boldsymbol{y}$ und $\hat{\boldsymbol{y}}$:

$$r^2_{\boldsymbol{y}\hat{\boldsymbol{y}}} = \frac{\left[\sum_{i=1}^m (y^{(i)} - \bar{y})(\hat{y}^{(i)} - \bar{\hat{y}})\right]^2}{\sum_{i=1}^m (y^{(i)} - \bar{y})^2 \sum_{i=1}^m (\hat{y}^{(i)} - \bar{\hat{y}})^2} = \frac{\left[\sum_{i=1}^m (y^{(i)} + e^{(i)} - \bar{y})(\hat{y}^{(i)} - \bar{\hat{y}})\right]^2}{\sum_{i=1}^m (y^{(i)} - \bar{y})^2 \sum_{i=1}^m (\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

$$= \frac{\left[\sum_{i=1}^m (\hat{y}^{(i)} - \bar{y})(\hat{y}^{(i)} - \bar{y})\right]^2}{\sum_{i=1}^m (y^{(i)} - \bar{y})^2 \sum_{i=1}^m (\hat{y}^{(i)} - \bar{\hat{y}})^2} = \frac{\sum_{i=1}^m (\hat{y}^{(i)} - \bar{y})^2}{\sum_{i=1}^m (y^{(i)} - \bar{y})^2} = \frac{ESS}{TSS}$$

$$= R^2 \ .$$

Bemerkung:
Im univariaten Fall, d.h. mit nur einer erklärenden Variable x, gilt zusätzlich

$$R^2 = r^2_{xy} \ .$$

Test auf Signifikanz des Gesamtmodells

Anpassungsgüte vs. Signifikanz

- Mit Hilfe des Bestimmtheitsmaßes lässt sich eine Aussage darüber treffen, wie gut ein Modell an die Daten anpasst.
- Dies bedeutet jedoch nicht, dass das Modell überhaupt richtig spezifiziert wurde (man kann immer eine lineare Funktion fitten, auch wenn der zu Grunde liegende Zusammenhang nichtlinear ist!).
- Man sollte daher einen Test auf Signifikanz des Modells durchführen.
- Der F-Test prüft die Güte der Vorhersage der Daten durch eine lineare Regressionsgleichung, indem untersucht wird, ob mindestens ein Steigungsparameter θ_i, $i \in \{1, \ldots, p\}$ sich signifikant von Null unterscheidet.

Test des Modells auf Signifikanz

Um Hypothesentests durchzuführen, brauchen wir ein probabilistisches Modell. Bei der linearen Regression werden üblicherweise folgende Modellannahmen gemacht:

Probabilistisches Modell für die multivariate lineare Regression

$$Y^{(i)} = \theta_0 + \theta_1 X_1^{(i)} + \ldots + \theta_p X_p^{(i)} + \varepsilon^{(i)} \ , \tag{1}$$

wobei die Störgrößen $\varepsilon^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, m$, unabhängig identisch normalverteilt sind.

- Die Größen $Y^{(i)}$ und $X_p^{(i)}$ und $\varepsilon^{(i)}$ repräsentieren Zufallsvariablen, die gemessenen Daten $(x^{(i)}, y^{(i)})$ werden als Realisierungen aufgefasst.
- Die anhand der Daten gefitteten Parameter $\hat{\boldsymbol{\theta}}$ repräsentieren eine Schätzung der Parameter $\theta_0, \ldots, \theta_p$ aus (1).

F-Test

Mit Hilfe des F-Tests kann untersucht werden, ob das Gesamtmodell signifikant ist, d.h., dass mindestens eine der unabhängigen Variablen einen statistisch signifikanten Einfluss auf die Zielvariable hat. Konkret wird die **Nullhypothese**

$$H_0 : \ \theta_1 = \theta_2 = \ldots = \theta_p = 0$$

gegen die **Alternativhypothese** $H_1 : \ \theta_j \neq 0$ für mindestens ein $j \in \{1, \ldots, p\}$ getestet. Die **Teststatistik** lautet

$$F = \frac{R^2}{1 - R^2} \cdot \frac{m - p - 1}{p} \ .$$

Sie ist unter Annahme der Nullhypothese F-verteilt mit p und $m - p - 1$ Freiheitsgraden. Die Nullhypothese wird verworfen, falls

$$F > F_{1-\alpha}(p, m - p - 1) \ ,$$

wobei $F_{1-\alpha}(p, m - p - 1)$ das Quantil der F-Verteilung mit p und $m - p - 1$ Freiheitsgraden zum Signifikanzniveau α ist.

Bemerkungen

- Selbst wenn man die Nullhypothese nicht verwerfen kann, gibt es noch Maßnahmen, die man ergreifen kann, z.B.
  - Entfernung von Ausreißern in den Daten
  - Einführung nichtlineare Transformationen der Variablen.
  - Hinzunahme weiterer Daten
- Wenn das Testergebnis dazu führt, dass die Nullhypothese nicht verworfen wird, kann man nicht folgern, dass tatsächlich kein linearer Zusammenhang besteht. Auch wenn man die Nullhypothese nicht verwerfen kann, ist es möglich, dass es einen linearen Zusammenhang zwischen den Variablen gibt. Es kann einfach sein, dass die Datenlage zu dünn ist, um einen tatsächlich vorhandenen Effekt statistisch zu belegen.

Beispiel

Gegeben seien die folgenden Daten:

x	1.0	2.0	3.0	2.5	4.0
y	2.0	4.0	2.0	3.0	1.0

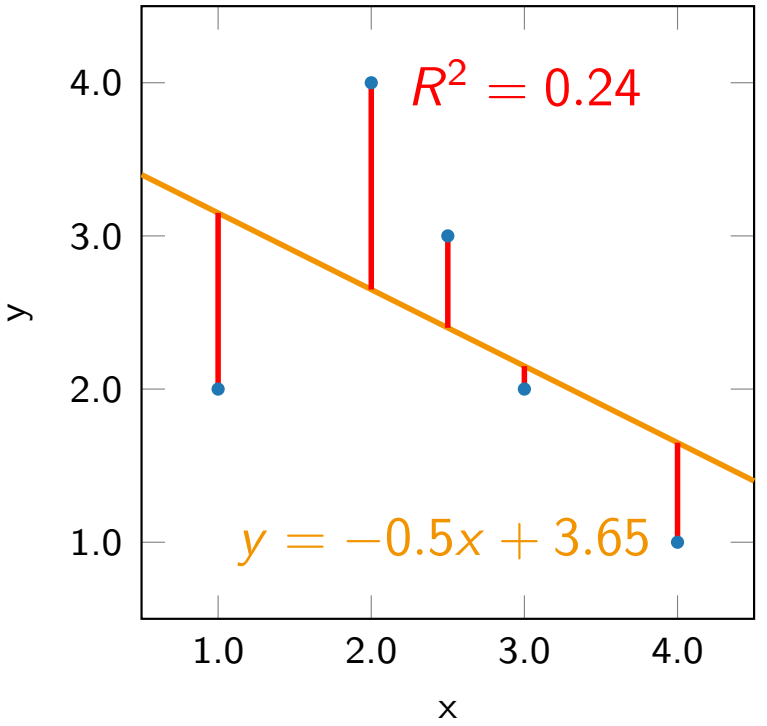
Es gilt also $m = 5$, $p = 1$. Die Regressionsgerade lautet $y = -0.5x + 3.65$.

Bestimmtheitsmaß

F-Statistik

Referenzstatistik

Testentscheidung:



Beispiel

Gegeben seien die folgenden Daten:

x	1.0	2.0	3.0	2.5	4.0
y	2.0	4.0	2.0	3.0	1.0

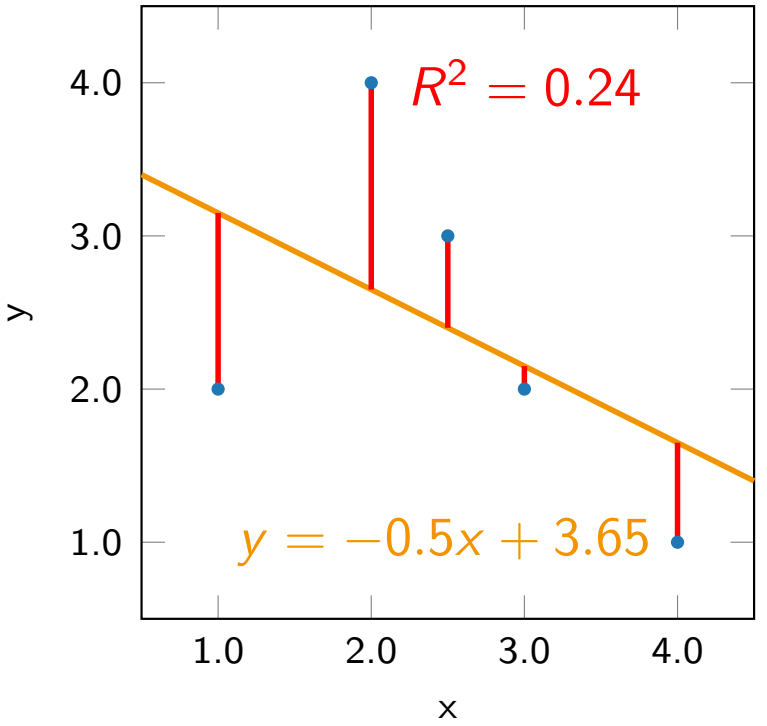
Es gilt also $m = 5$, $p = 1$. Die Regressionsgerade lautet $y = -0.5x + 3.65$.

Bestimmtheitsmaß
$R^2 = 0.24$

F-Statistik

Referenzstatistik

Testentscheidung:



Beispiel

Gegeben seien die folgenden Daten:

x	1.0	2.0	3.0	2.5	4.0
y	2.0	4.0	2.0	3.0	1.0

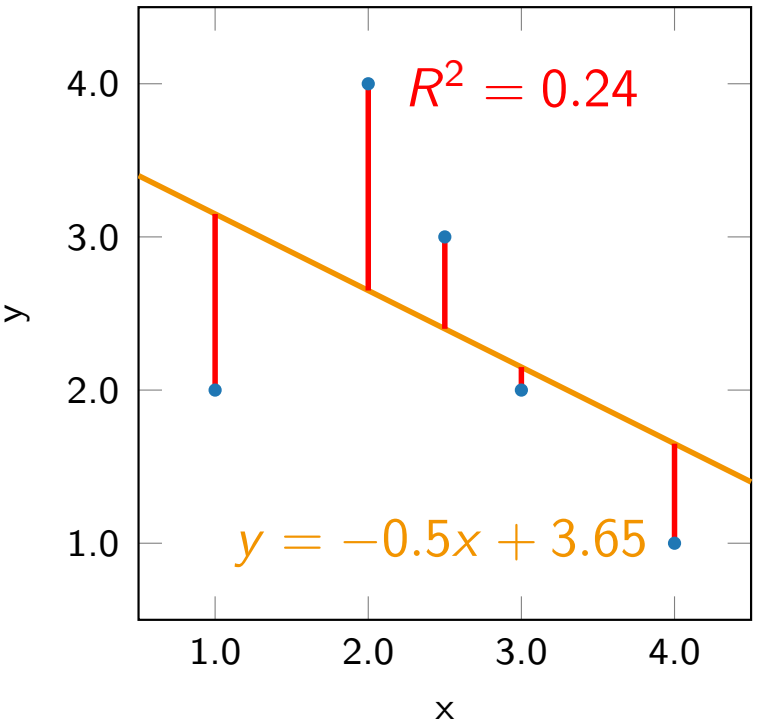
Es gilt also $m = 5$, $p = 1$. Die Regressionsgerade lautet $y = -0.5x + 3.65$.

Bestimmtheitsmaß
$R^2 = 0.24$

F-Statistik
$F = \frac{R^2}{1-R^2} \cdot \frac{m-p-1}{p} = \frac{0.24}{0.76} \cdot \frac{3}{1} = 0.949$

Referenzstatistik

Testentscheidung:



Beispiel

Gegeben seien die folgenden Daten:

x	1.0	2.0	3.0	2.5	4.0
y	2.0	4.0	2.0	3.0	1.0

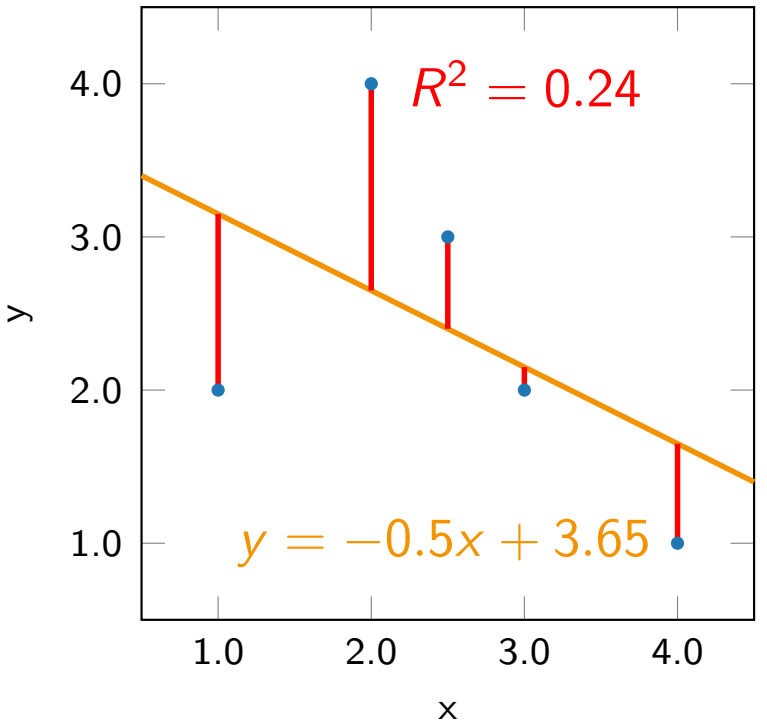
Es gilt also $m = 5$, $p = 1$. Die Regressionsgerade lautet $y = -0.5x + 3.65$.

Bestimmtheitsmaß
$R^2 = 0.24$

F-Statistik
$F = \frac{R^2}{1-R^2} \cdot \frac{m-p-1}{p} = \frac{0.24}{0.76} \cdot \frac{3}{1} = 0.949$

Referenzstatistik
$F_{1-\alpha}(p, m-p-1) = F_{0.95}(1,3)$
$= 10.128.$ Tabelle

Testentscheidung:



Beispiel

Gegeben seien die folgenden Daten:

x	1.0	2.0	3.0	2.5	4.0
y	2.0	4.0	2.0	3.0	1.0

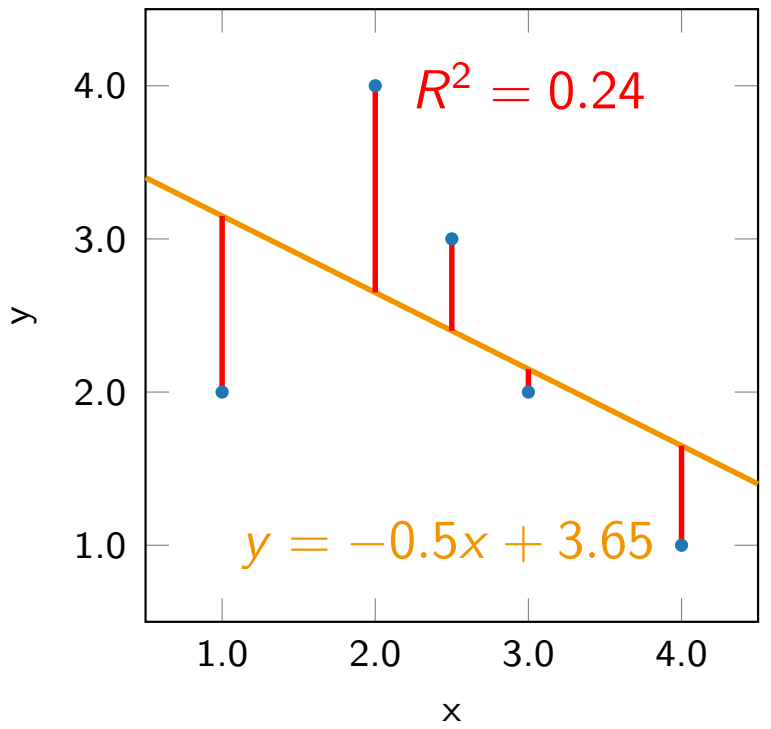
Es gilt also $m = 5$, $p = 1$. Die Regressionsgerade lautet $y = -0.5x + 3.65$.

Bestimmtheitsmaß
$R^2 = 0.24$

F-Statistik
$F = \frac{R^2}{1-R^2} \cdot \frac{m-p-1}{p} = \frac{0.24}{0.76} \cdot \frac{3}{1} = 0.949$

Referenzstatistik
$F_{1-\alpha}(p, m-p-1) = F_{0.95}(1,3)$
$= 10.128.$

Testentscheidung:
$F < F_{1-\alpha}(p, m-p-1)$
$\Rightarrow$ H_0 wird zum Niveau $\alpha = 0.05$ nicht verworfen.



Quantile der F-Verteilung

Die Quantile der F-Verteilung sind (pro α) tabelliert. Für unser Beispiel liest man den Wert 10.13 aus der Tabelle ab:

Quantile der F-Verteilung $F_{0.95}(n_1, n_2)$

$n_2 \backslash n_1$	1	2	3	4	5	6
1	161.45	199.50	215.71	224.58	230.16	233.99
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.10	3.71	3.48	3.33	3.22

Zusammenfassung Bestimmtheitsmaß und F-Test

Vorteile:

- Dimensionslose Größe
- Wertebereich zwischen 0 und 1
- gute Interpretierbarkeit

Grenzen:

- Das Bestimmtheitsmaß gibt Auskunft über die Qualität der Approximation, aber nicht, ob das Modell korrekt spezifiziert wurde. Beispielsweise kann ein Modell mit hohem R^2 eine hohe Verzerrung haben.
- Ein hoher R^2 ist notwendig, aber nicht hinreichend für ein Modell mit guter Vorhersagekraft.
- Das Bestimmtheitsmaß sagt nichts über die statistische Signifikanz des ermittelten Zusammenhangs und der einzelnen erklärenden Variablen aus. Dazu kann ein F-Test durchgeführt werden.
- Das Bestimmtheitsmaß macht keine Aussage über einen kausalen Zusammenhang zwischen den Ein- und Ausgabegrößen.