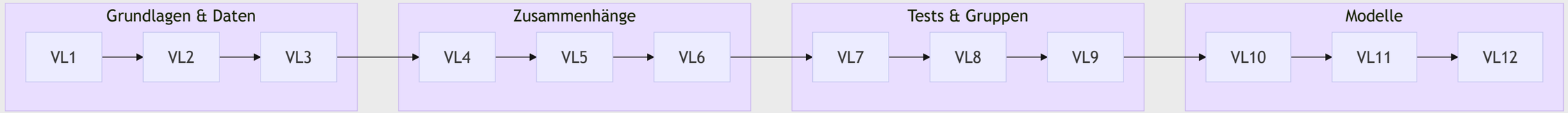
Statistik für Data Scientists

Vorlesung 10: Regression

Prof. Dr. Siegfried Handschuh
DS-NLP
Universität St. Gallen

Lernreise

- VL1–VL3: saubere Datenbasis schaffen
  Ihr lernt, Daten zu verstehen, zu säubern und sinnvoll zu visualisieren.
  Das ist die Grundlage für jede Analyse (z. B. WHO, OECD, NYC Parking).
- VL4–VL6: Zusammenhänge erkennen & Unsicherheit quantifizieren
  Korrelation, Zufall, Verteilungen, Konfidenzintervalle.
  Ohne dieses Fundament kann man Projektergebnisse nicht korrekt interpretieren.
- VL7–VL9: Hypothesen testen & Gruppen vergleichen
  Ihr beantwortet konkrete Forschungsfragen aus euren Datensätzen:
  «Unterscheiden sich Gruppen? Wie gross ist der Effekt? Sind die Ergebnisse robust?»
- Ab VL10: Modelle bauen, nicht nur testen
  Mit Regression erklärt ihr mehrere Einflüsse gleichzeitig.
  Genau das braucht ihr für die Projekte, bspw. WHO-Lebenserwartung, OECD-Wellbeing, SBB-Daten, etc.
- VL11: Realistische Modelle für reale Daten
  Confounding, Multikollinearität, Modellvergleich – Themen, die in euren Projekten immer auftauchen.
- VL12: Alles zusammenführen
  Ihr verbindet Statistik, Modellierung und Diagnoseplots zu einem vollständigen Analyseworkflow.


Grundlagen & Daten
VL1
VL2
VL3
Zusammenhänge
VL4
VL5
VL6
Tests & Gruppen
VL7
VL8
VL9
Modelle
VL10
VL11
VL12

Lernziele heute

Nach dieser Vorlesung kannst du:

- erklären, was eine einfache lineare Regression modelliert
- β_0 und β_1 interpretieren
- Residuen und Fehlerbegriffe verstehen
- die Grössen RSS, TSS, ESS und R^2 berechnen und interpretieren
- Standardfehler der Schätzung einordnen
- Diagnoseplots (Residuen, QQ, Cook's Distance) anwenden
- Regression sinnvoll in Projekten einsetzen

Notation in der Regression

- y_i: beobachteter Wert
- x_i: Prädiktorwert
- $\bar{y}$: Mittelwert aller y_i
- $\hat{y}_i$: vorhergesagter Wert des Modells
- $e_i = y_i - \hat{y}_i$: Residuum
- β_0: Achsenabschnitt
- β_1: Steigung
- TSS, RSS, ESS: Variationszerlegung

Einstieg

Warum Regression?

Regression beschreibt, wie sich Y verändert, wenn sich X verändert.

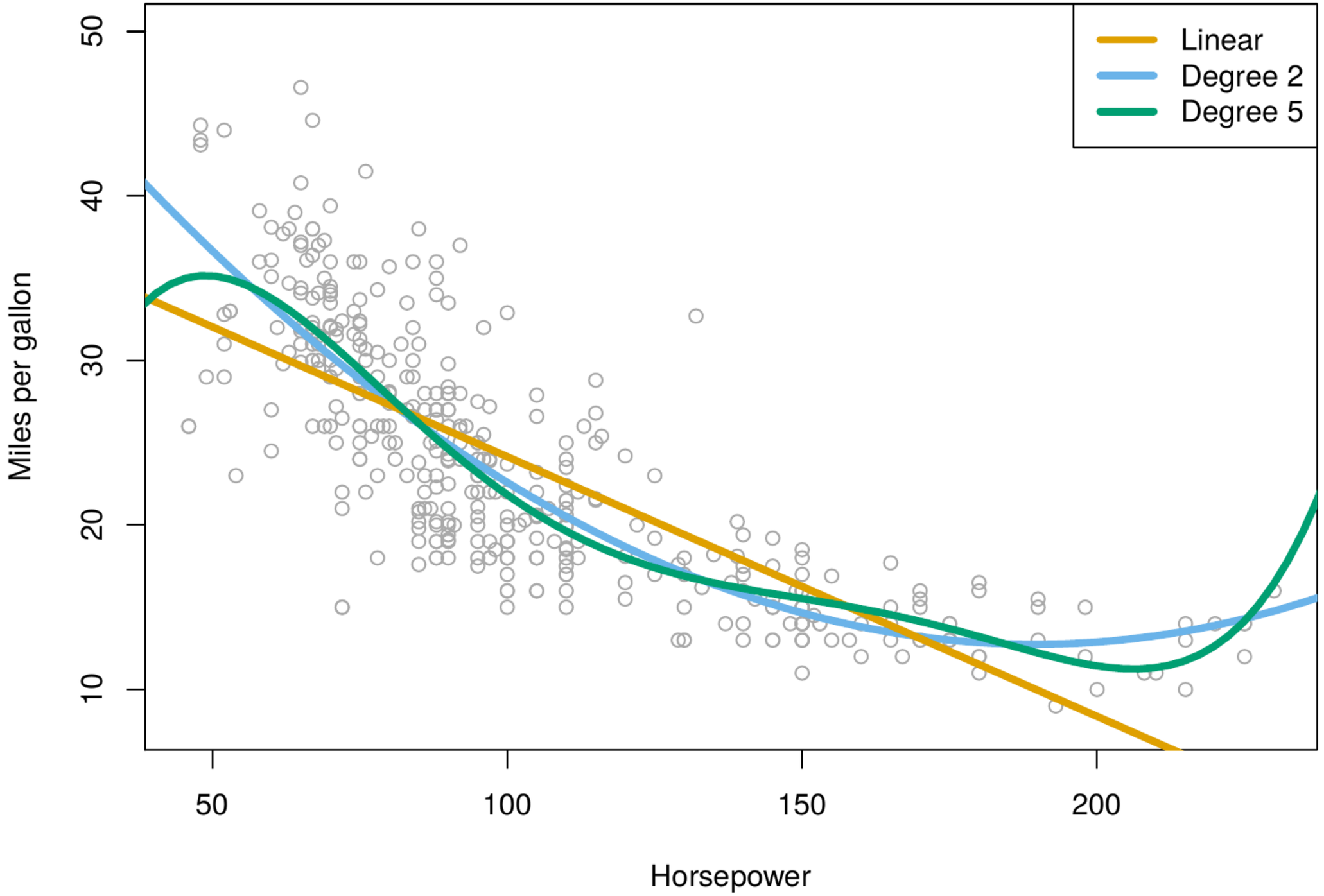
- Frage: Wie stark und in welche Richtung wirkt X auf Y?
- Beispiele:
  - WHO: Mortalität → Lebenserwartung
  - IMF: Inflation → Arbeitslosenquote
  - NYC Parking: Bussen → Tageszeit
  - SBB: Verspätung → Tageszeit
- Regression liefert eine gerichtete quantitative Beziehung.

Mini-Check: Was ist der Unterschied zwischen einer Veränderung in X und einer Korrelation mit X?

Lineares Modell als Approximation

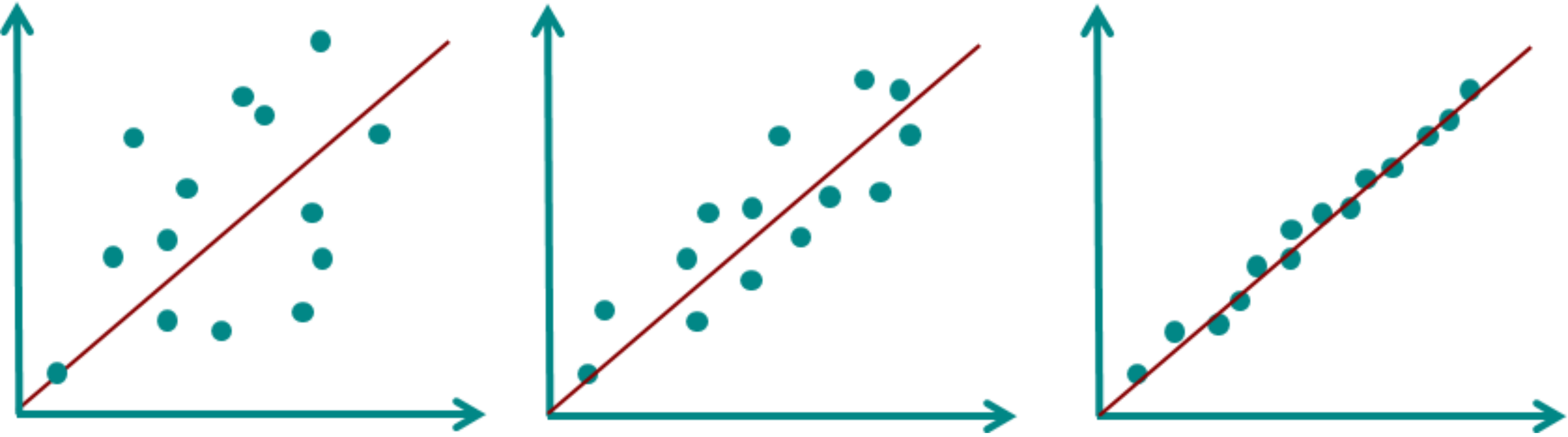
- Reale Zusammenhänge sind selten perfekt linear.
- Approximiert die Beziehung zwischen X und Y.
- Ziel: einfache, interpretierbare Funktion als Startpunkt.

Mini-Check: Warum verwenden wir trotz Nichtlinearität oft eine lineare Approximation?


Linear
Degree 2
Degree 5
Miles per gallon
10
20
30
40
50
50
100
150
200
Horsepower

Linearer Zusammenhang

Gering → Hoch

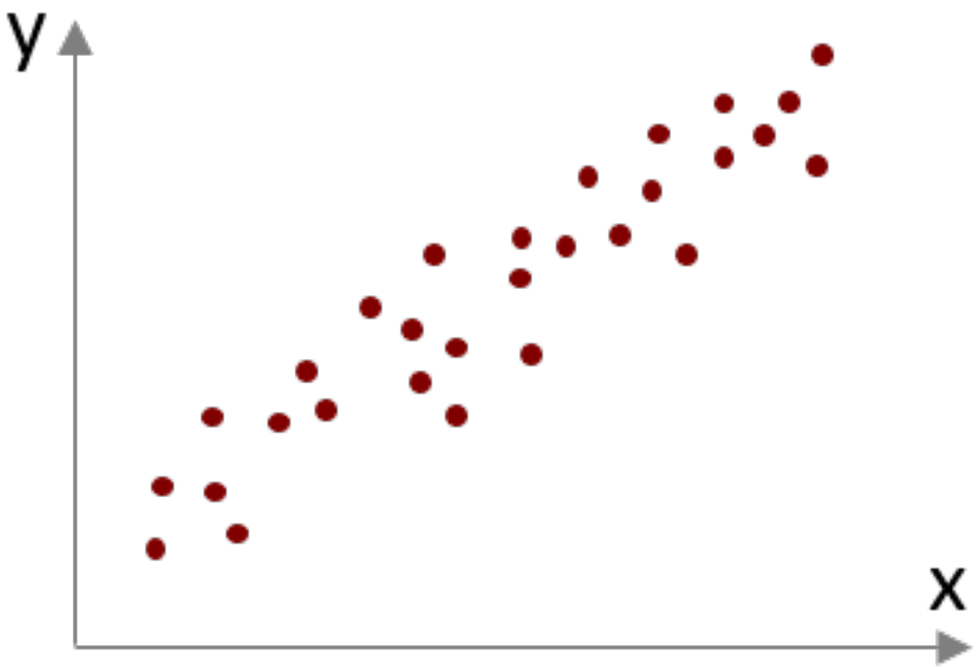


Regression vs Korrelation

Regression hat eine Richtung, Korrelation nicht.

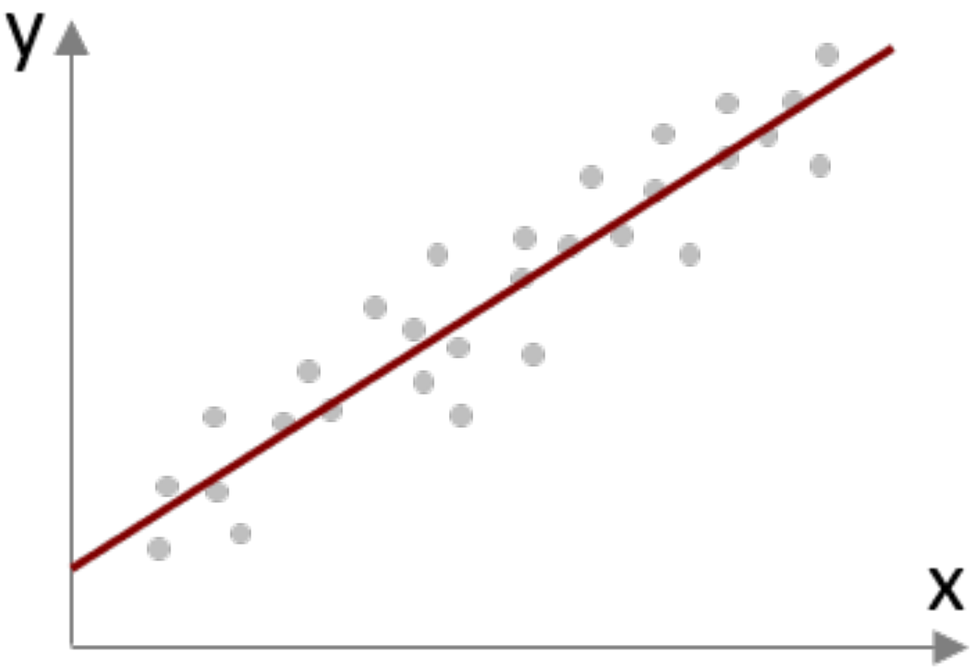
- Regression beantwortet die Frage: Wie verändert sich Y, wenn sich X verändert?
- Korrelation beschreibt nur, wie stark zwei Variablen gemeinsam variieren.
- Regression modelliert einen gerichteten Zusammenhang $X \rightarrow Y$.
- Korrelation ist symmetrisch: $\rho(X, Y) = \rho(Y, X)$.
- Modellform der Regression: $Y = \beta_0 + \beta_1 X + \epsilon$.

Mini-Check: Was liefert eine Richtung: Regression oder Korrelation?



r = .8

Korrelation



y = 1 + 0,5x

Regression

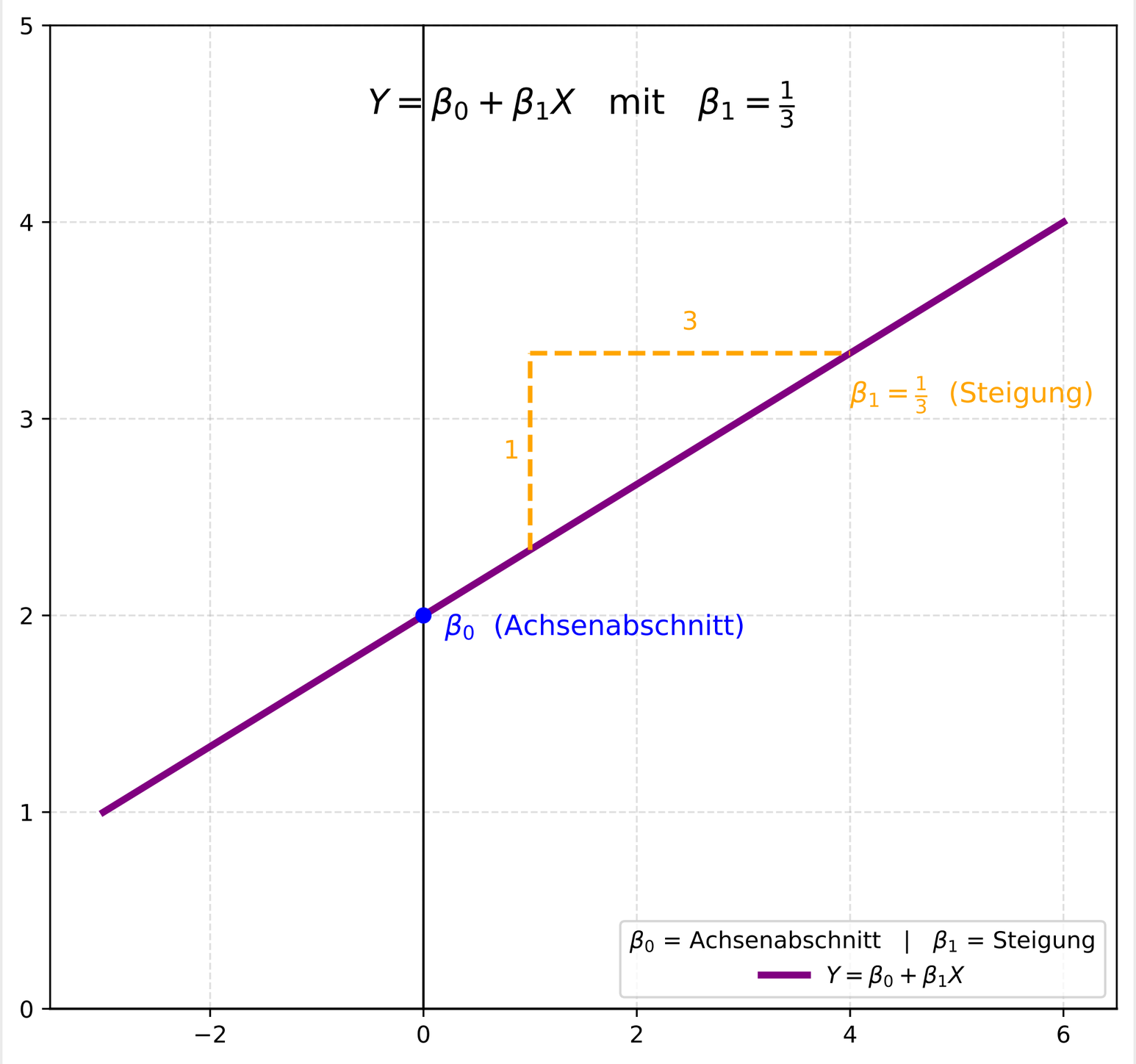
Einfache Regression

Lineares Modell

Die einfache lineare Regression beschreibt Y als lineare Funktion von X plus Fehler.

- Modellidee: $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 ist der Achsenabschnitt: erwartetes Y, wenn $X = 0$.
- β_1 ist die Steigung: Veränderung von Y pro Einheit X.
- ϵ repräsentiert alle Einflüsse, die das Modell nicht erklärt.
- Schätzung erfolgt mittels OLS: Minimierung der Summe der quadrierten Fehler.

Mini-Check: Was bedeutet ein positiver β_1?


$Y = \beta_0 + \beta_1 X$ mit $\beta_1 = \frac{1}{3}$
3
1
$\beta_1 = \frac{1}{3}$ (Steigung)
β_0 (Achsenabschnitt)
β_0 = Achsenabschnitt | β_1 = Steigung
$Y = \beta_0 + \beta_1 X$
5
4
3
2
1
0
−2
0
2
4
6

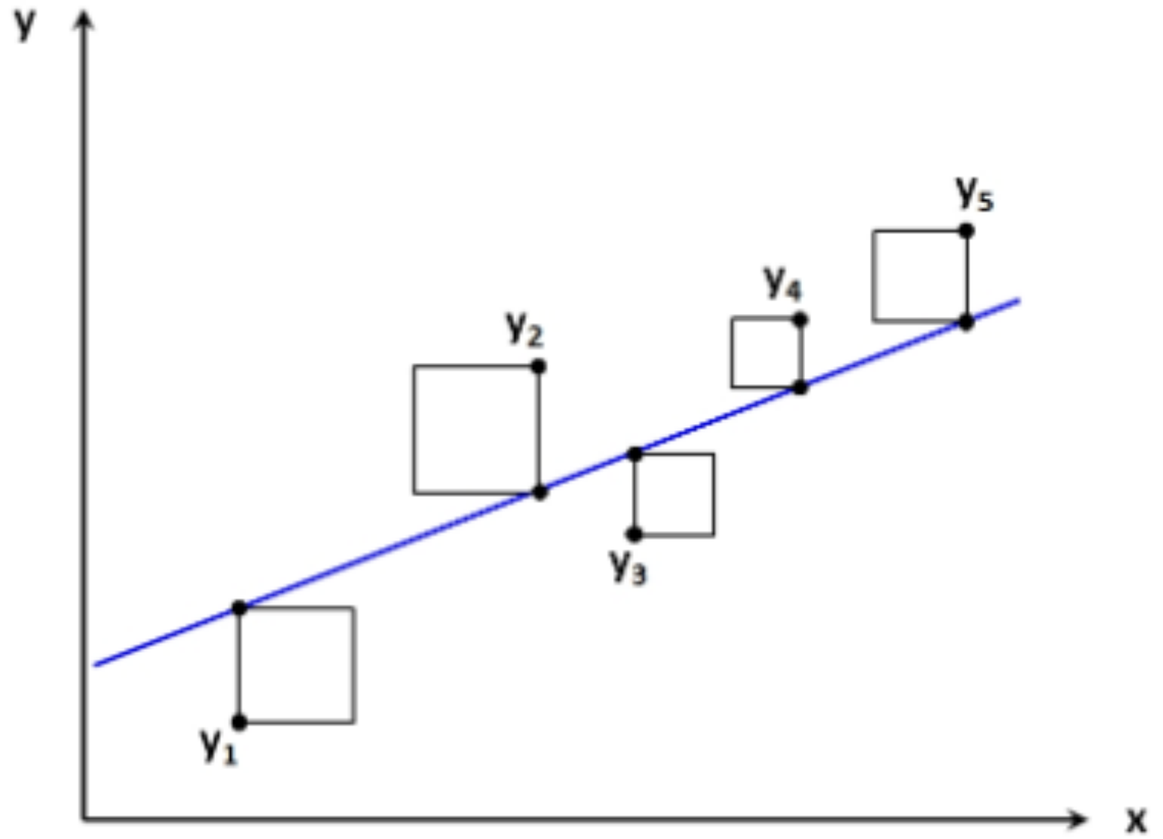
Ordinary Least Squares (OLS)

Die Regressionsgerade wird so gewählt, dass die Quadrierte Fehler-Summe minimal wird.

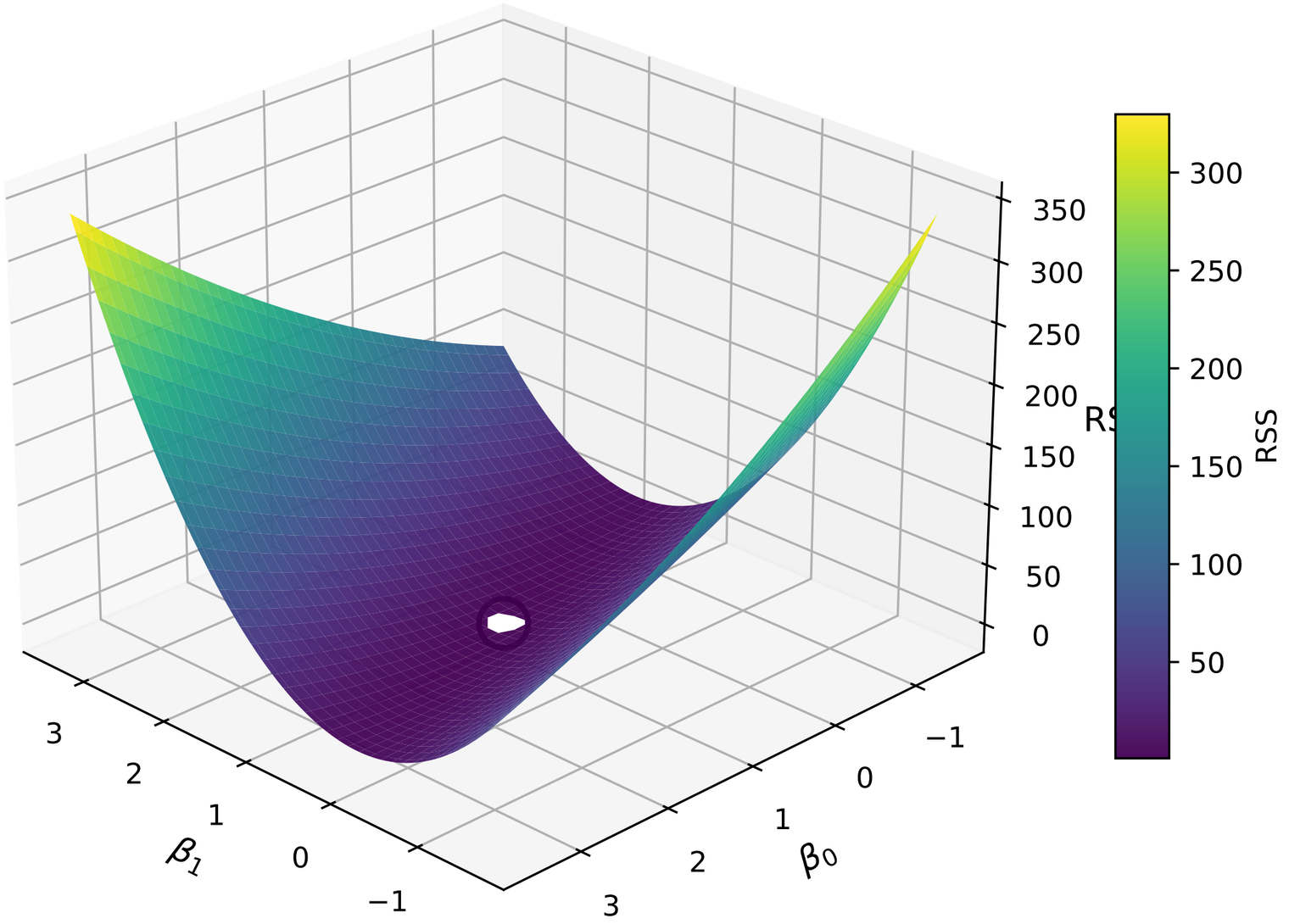
- Residuen: $e_i = y_i - \hat{y}_i$
- OLS minimiert die Residual Sum of Squares:
  $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- grosse Abweichungen werden stärker bestraft
- ergibt eine eindeutige, effizient berechenbare Lösung

Mini-Check: Warum quadrieren wir die Fehler statt sie absolut zu nehmen?

Methode der kleinsten Quadrate



RSS-Oberfläche mit Loch am OLS-Minimum
Weisser Punkt = OLS-Lösung



Mini-Beispiel: Lineares Modell mit Zahlen

Wir betrachten einfache Daten:

- $X = [1, 2, 3, 4]$
- $Y = [2, 3, 3, 5]$

Durch OLS ergibt sich:

- $\hat{y} = 1.5 + 0.8x$

Beispiele:

- Für $X = 2 \rightarrow \hat{Y} = 3.1$
- Für $X = 5 \rightarrow \hat{Y} = 5.5$

Mini-Check: Was bedeutet $\beta_1 = 0.8$ in diesem Kontext?

Python: Statsmodels

```
import statsmodels.api as sm
import numpy as np

X = np.array([1,2,3,4])
Y = np.array([2,3,3,5])

X = sm.add_constant(X)      # fügt β0 hinzu
model = sm.OLS(Y, X).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	11.57
Date:	Sun, 30 Nov 2025	Prob (F-statistic):	0.0766
Time:	17:56:29	Log-Likelihood:	-2.1898
No. Observations:	4	AIC:	8.380
Df Residuals:	2	BIC:	7.152
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0000	0.725	1.380	0.302	-2.118	4.118
x1	0.9000	0.265	3.402	0.077	-0.238	2.038

Omnibus:	nan	Durbin-Watson:	2.900
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.678
Skew:	-0.922	Prob(JB):	0.712
Kurtosis:	2.183	Cond. No.	7.47

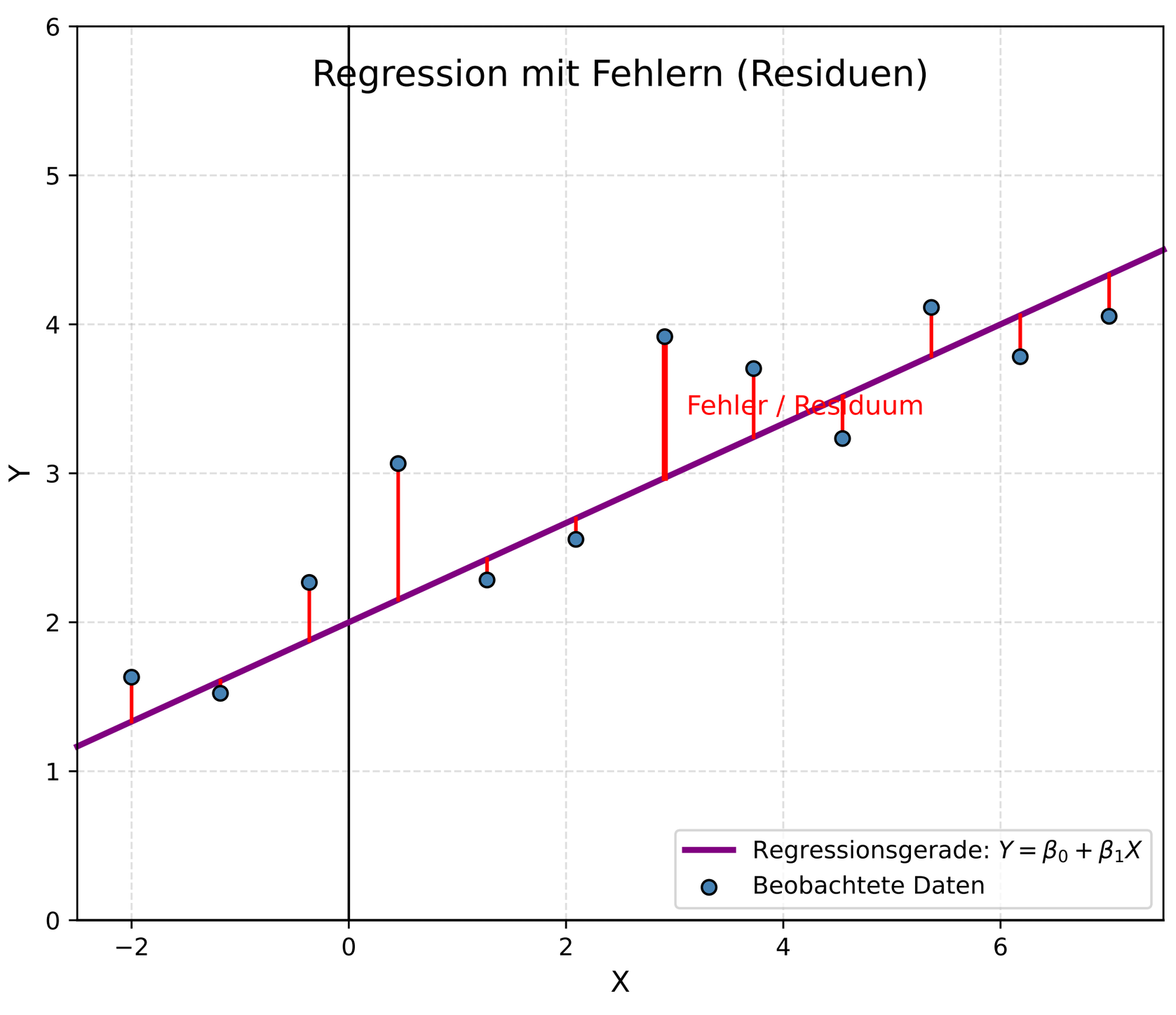
Python: scikit-learn

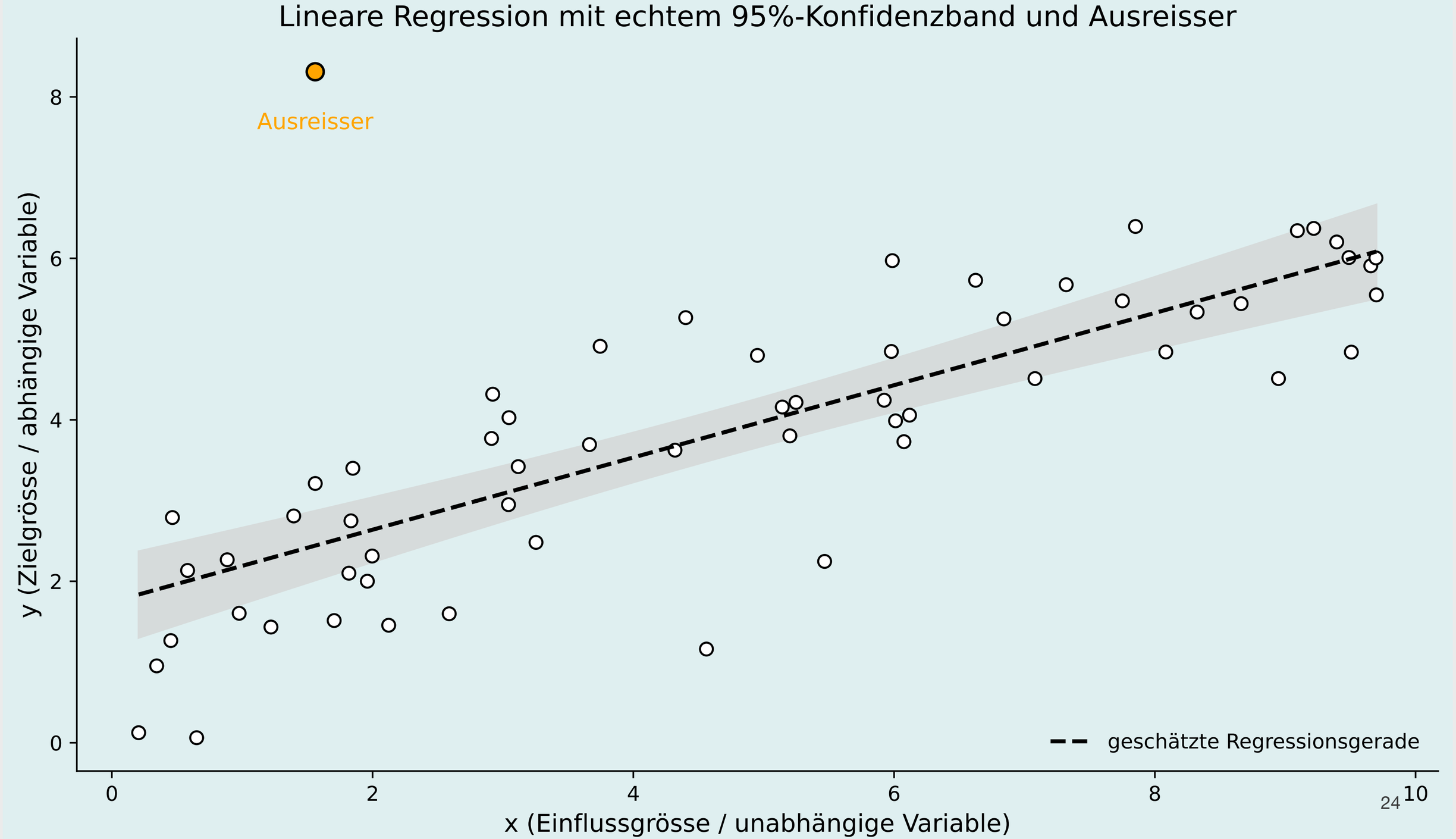
```
from sklearn.linear_model import LinearRegression
import numpy as np

# X muss eine 2D-Matrix sein → reshape
X = np.array([1,2,3,4]).reshape(-1,1)
Y = np.array([2,3,3,5])

model = LinearRegression().fit(X,Y)
print(model.intercept_, model.coef_)
```

```
1.0000000000000009 [0.9]
```


Regression mit Fehlern (Residuen)
Fehler / Residuum
Regressionsgerade: $Y = \beta_0 + \beta_1 X$
Beobachtete Daten
X
Y
0
1
2
3
4
5
6
−2
0
2
4
6


Lineare Regression mit echtem 95%-Konfidenzband und Ausreisser
Ausreisser
y (Zielgrösse / abhängige Variable)
x (Einflussgrösse / unabhängige Variable)
geschätzte Regressionsgerade
0
2
4
6
8
10

Residuen

Residuen sind die Differenz zwischen beobachtetem und vorhergesagtem Wert.

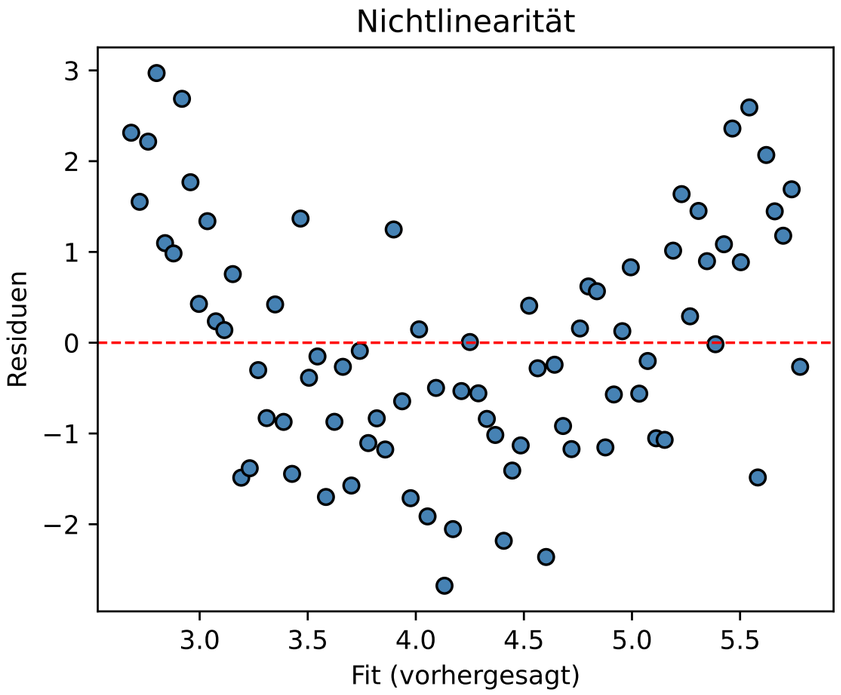
Residuum = beobachtetes Y minus vorhergesagtes $\hat{Y}$

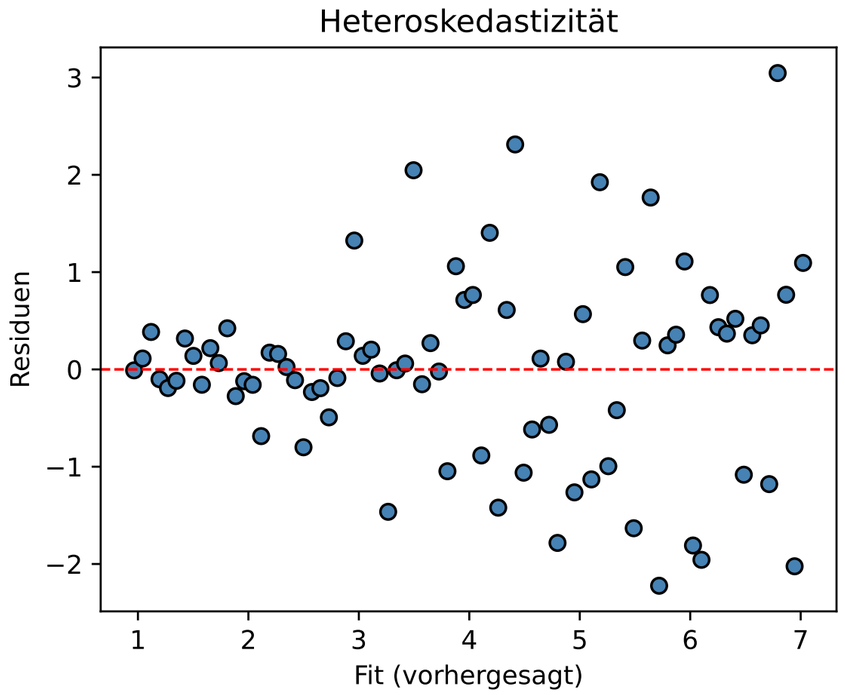
grosse Residuen → Modell passt schlecht

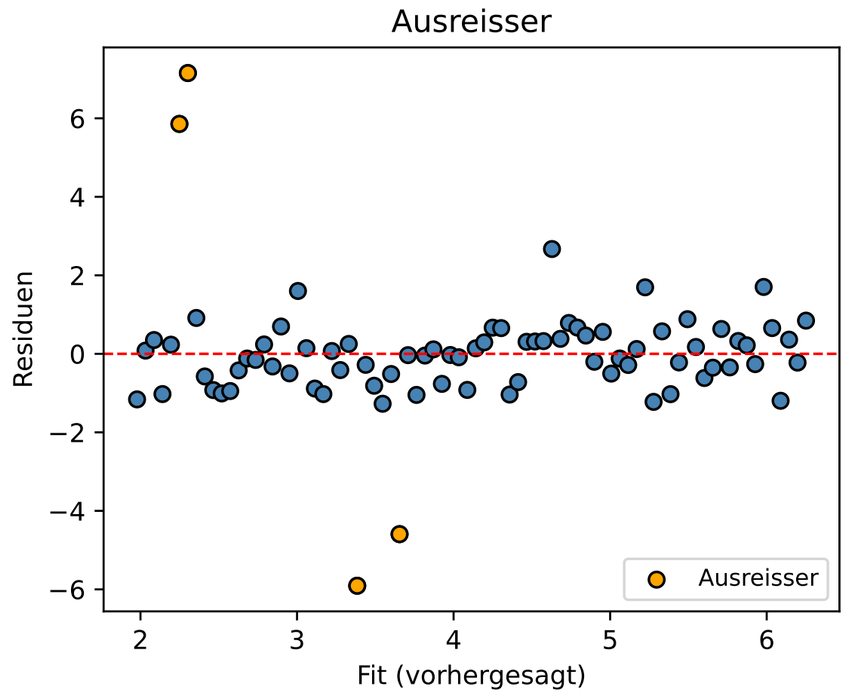
Muster in Residuen zeigen Probleme:

- Nichtlinearität
- Heteroskedastizität
- Ausreisser

Ziel: Residuen sollen «zufällig» aussehen


Nichtlinearität
Residuen
Fit (vorhergesagt)
3
2
1
0
−1
−2
3.0
3.5
4.0
4.5
5.0
5.5


Heteroskedastizität
Residuen
Fit (vorhergesagt)
3
2
1
0
−1
−2
1
2
3
4
5
6
7


Ausreisser
Residuen
Fit (vorhergesagt)
Ausreisser
6
4
2
0
−2
−4
−6
2
3
4
5
6

Gütemasse der Regression

Warum brauchen wir Gütemasse?

- Wir kennen jetzt die Regressionsgerade.
- Offene Frage: Wie gut beschreibt sie die Daten?
- Drei Perspektiven:
  - Wie viel Variation wird erklärt?
  - Wie gross sind die Fehler?
  - Wie verhalten sich die Residuen?

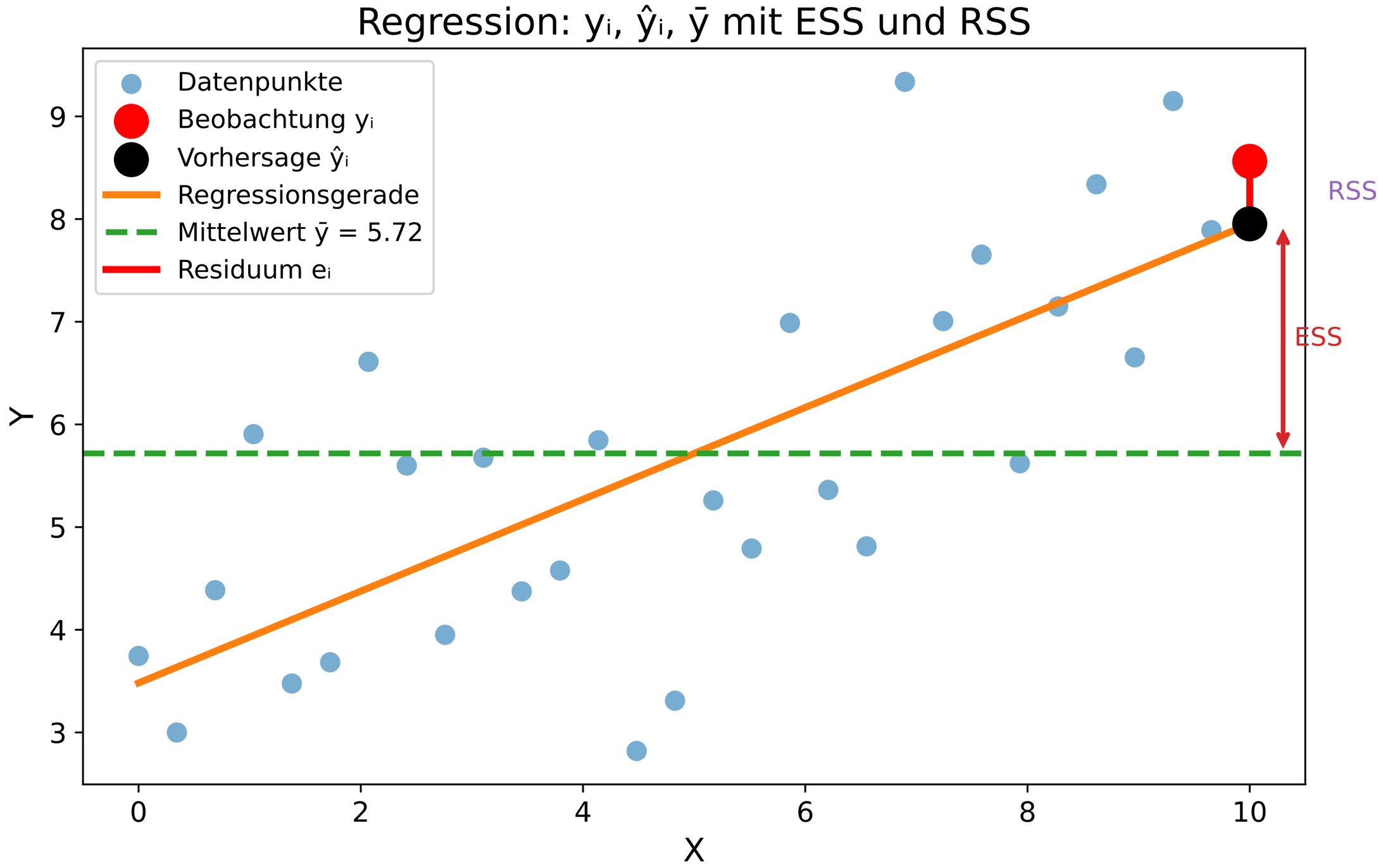
Mini-Check: Was wäre ein Anzeichen für ein «schlechtes» Modell?

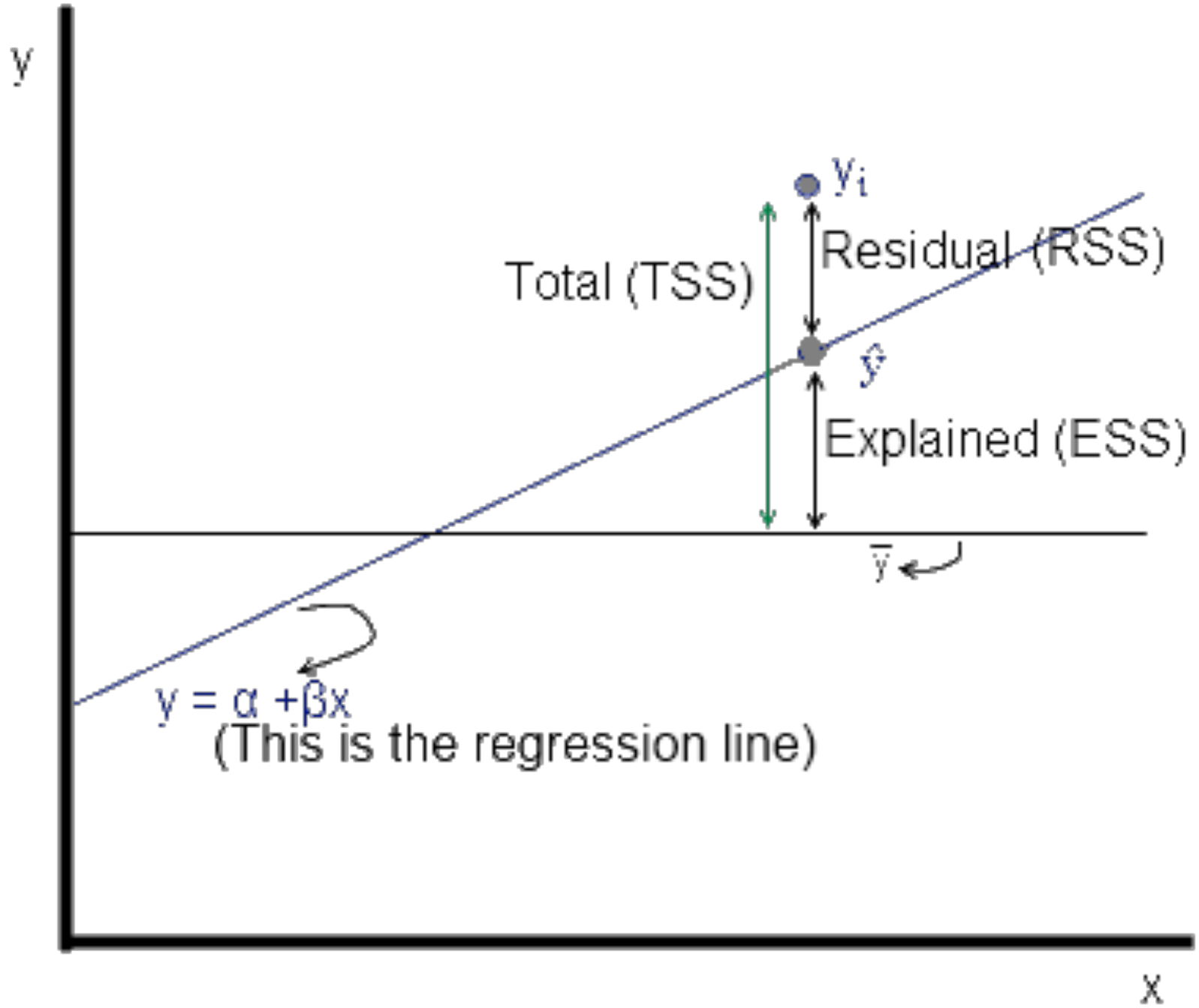
Zerlegung der Variation

Die totale Variation in Y lässt sich zerlegen in erklärte und unerklärte Anteile.

- Totale Variation: $TSS = \sum (y_i - \bar{y})^2$
- Unerklärte Variation: $RSS = \sum (y_i - \hat{y}_i)^2$
- Erklärte Variation: $ESS = TSS - RSS$
- Basis für Gütemasse wie R^2

Mini-Check: Was bedeutet ein kleines RSS?


Regression: yᵢ, ŷᵢ, ȳ mit ESS und RSS
Datenpunkte
Beobachtung yᵢ
Vorhersage ŷᵢ
Regressionsgerade
Mittelwert ȳ = 5.72
Residuum eᵢ
RSS
ESS
Y
X
0
2
4
6
8
10
3
4
5
6
7
8
9


y
y_i
Residual (RSS)
Total (TSS)
$\hat{y}$
Explained (ESS)
$\bar{y}$
$y = \alpha + \beta x$
(This is the regression line)
x

Bestimmtheitsmass R^2

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

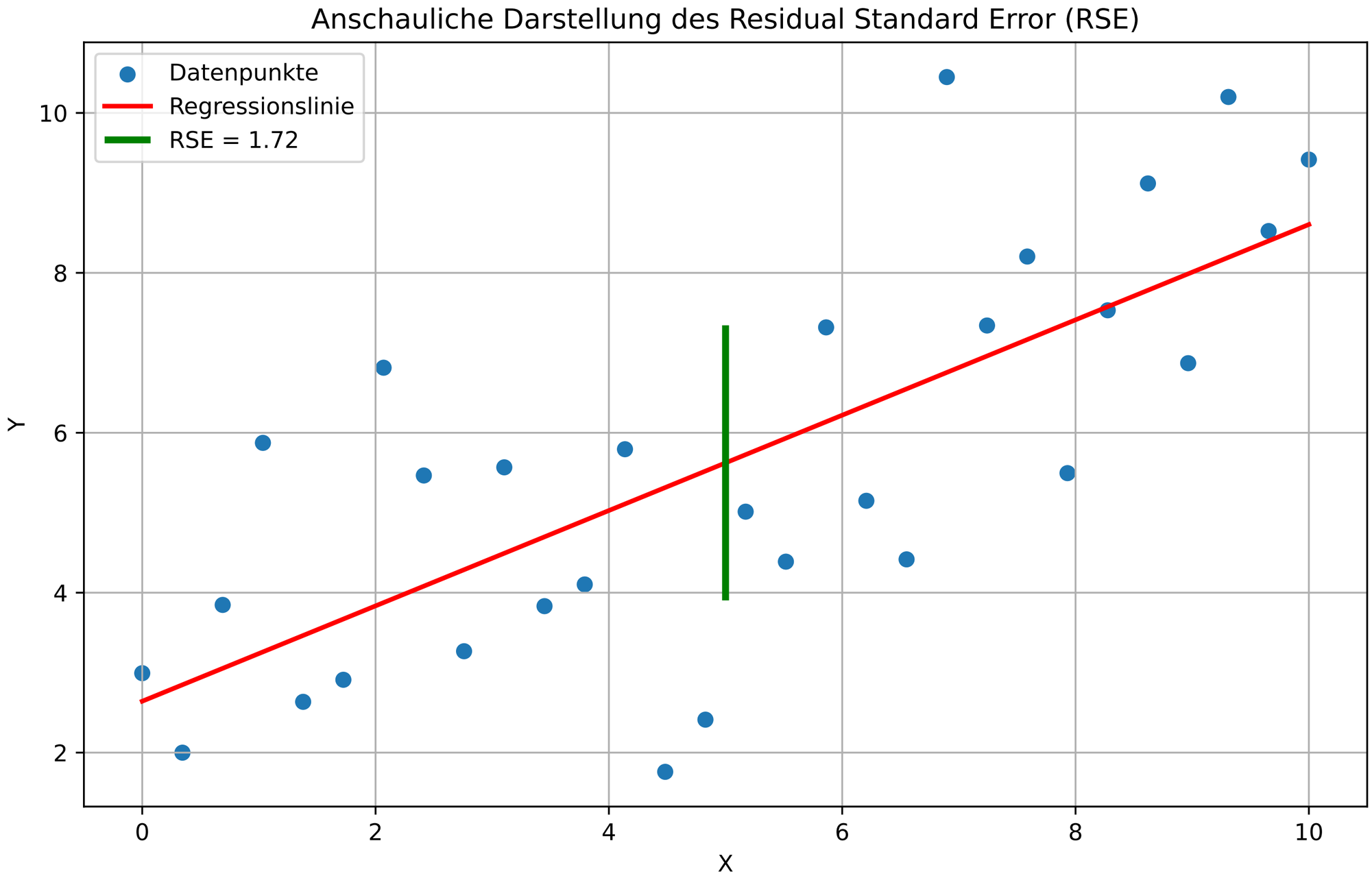
- misst den Anteil der erklärten Variation
- Wertebereich 0 bis 1
- höher ≠ immer besser (Overfitting möglich)
- Vorsicht: steigt auch ohne echte Verbesserung

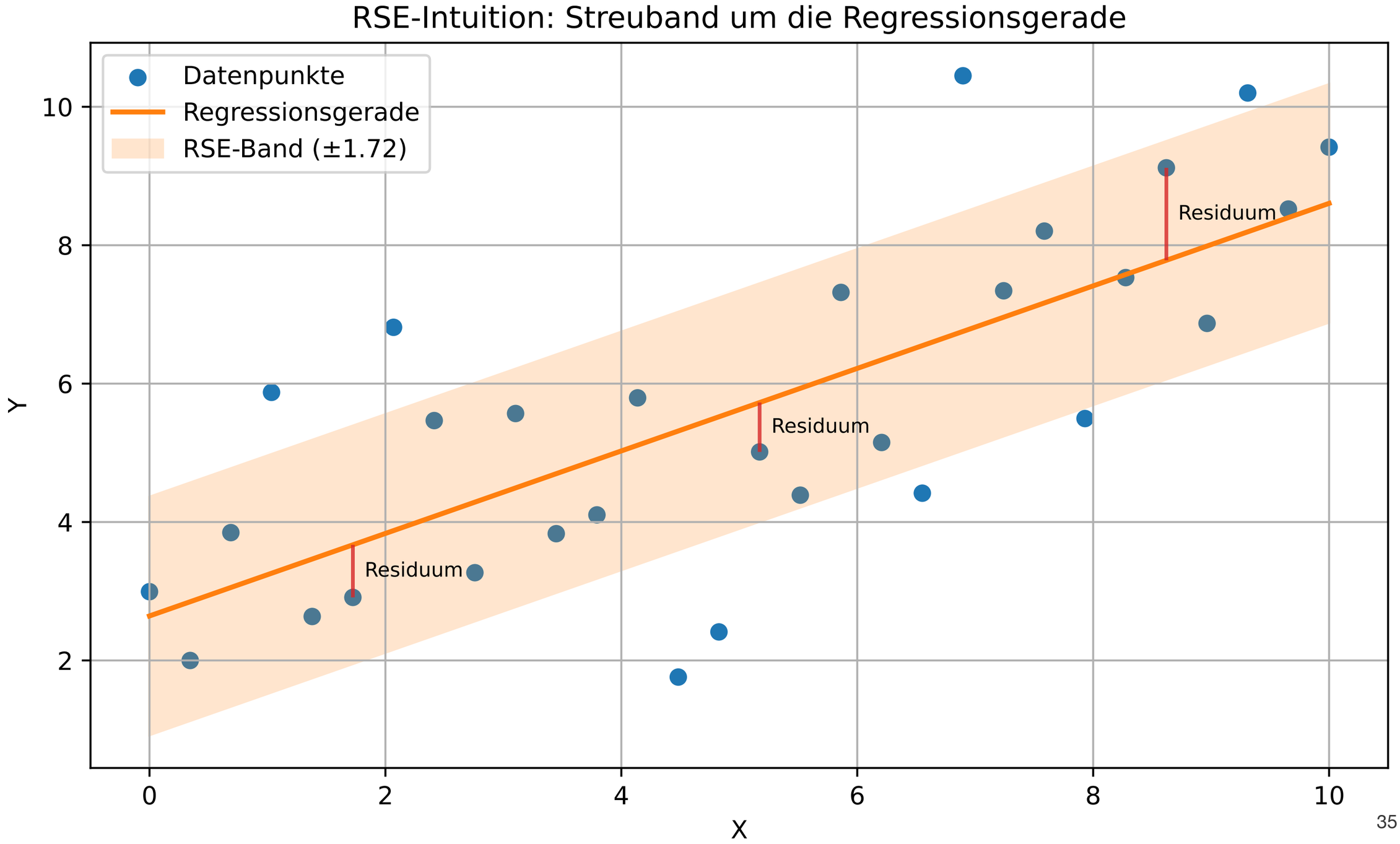
Mini-Check: Was bedeutet $R^2 = 0.8$?

Intuition: Was misst RSE?

- RSE = typische Abweichung eines Punkts von der Geraden.
- Je kleiner der RSE, desto enger liegen die Daten um die Gerade.
- RSE berücksichtigt die Freiheitsgrade (n – 2).

Mini-Check: Was sagt ein hoher RSE aus?


Anschauliche Darstellung des Residual Standard Error (RSE)
Datenpunkte
Regressionslinie
RSE = 1.72
Y
X
0
2
4
6
8
10
2
4
6
8
10


RSE-Intuition: Streuband um die Regressionsgerade
Datenpunkte
Regressionsgerade
RSE-Band (±1.72)
Residuum
Residuum
Residuum
Y
X
0
2
4
6
8
10
2
4
6
8
10

Residual Standard Error (RSE)

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

- misst durchschnittliche Abweichung der Residuen
- kleiner RSE → bessere Passung
- hängt von Massstab von Y ab
- Basis für Standardfehler der Koeffizienten

Mini-Check: Warum teilen wir durch $n-2$?

Erklärte Variation: ESS

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- misst, wie viel Variation das Modell erklärt
- je grösser ESS → desto informativer das Modell
- Gesamtvariation zerfällt in
  $TSS = ESS + RSS$
- Grundlage für R^2

Mini-Check: Was bedeutet ESS = 0?

Projektbezug: R² in realen Datensätzen

Beispiel WHO (Lebenserwartung ~ Mortalität):

- Modell erklärt z. B. $R^2 = 0.62$ der Variation.
- Interpretation: 62 % der Unterschiede in der Lebenserwartung werden durch Mortalität erklärt.
- Frage: Ist das für die Fragestellung ausreichend?

Mini-Check: Wäre $R^2 = 0.20$ auch akzeptabel?

Richtwerte

Anwendungsgebiet	Typische R^2-Werte	Was gilt als "gut"?
Physik / Ingenieurwesen	0.90 – 0.99	> 0.95
Medizinische Messungen (Labor)	0.70 – 0.95	> 0.85
Finanzmodelle (Makro)	0.30 – 0.70	> 0.50
Ökonometrie / Sozialwissenschaften	0.10 – 0.40	> 0.30
Marketing / Nutzerverhalten	0.05 – 0.30	> 0.20
Echte „Noisy Real-World Data“ (z. B. Städte, Mobilität)	0.00 – 0.30	> 0.20
Machine Learning (komplexe Modelle)	0.40 – 0.95	> 0.60 (modellspezifisch)
Time-Series Forecasting (kurzfristig)	0.20 – 0.70	> 0.50
Biologie / Genetik	0.05 – 0.20	> 0.15

Zwischenfazit Regression

Was wir bis hierhin aufgebaut haben:

- Lineares Modell als einfachste Funktionsannahme
- OLS zur Schätzung von β_0 und β_1
- Residuen als zentrales Diagnosewerkzeug
- Zerlegung der Varianz in TSS, ESS, RSS
- Gütekennzahlen wie R^2 und RSE zur Modellbewertung

Mini-Check: Welche zwei Grössen brauchst du, um R^2 zu berechnen?

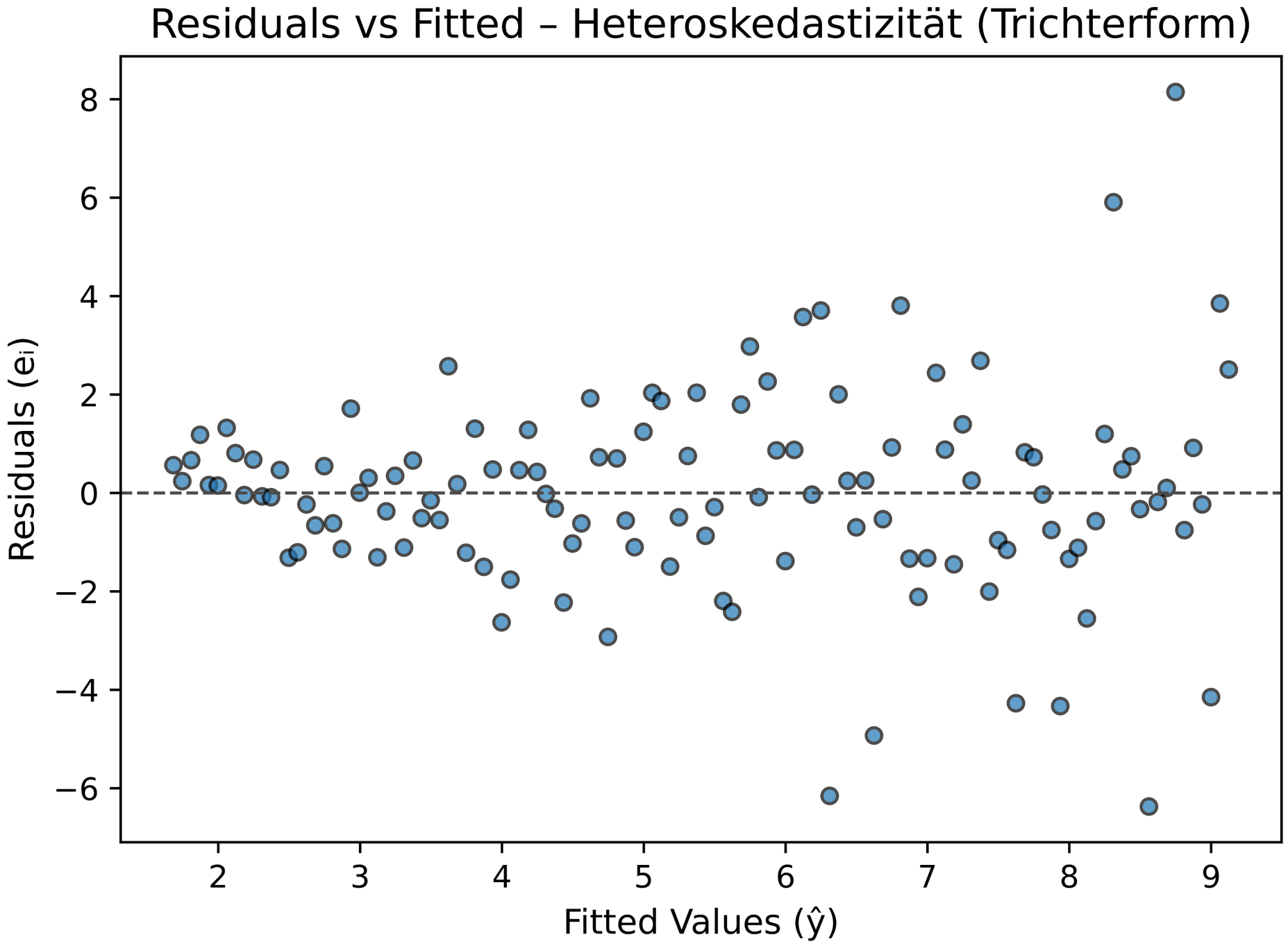
Residuenplots & Modellannahmen

Residuenplots & Modellannahmen

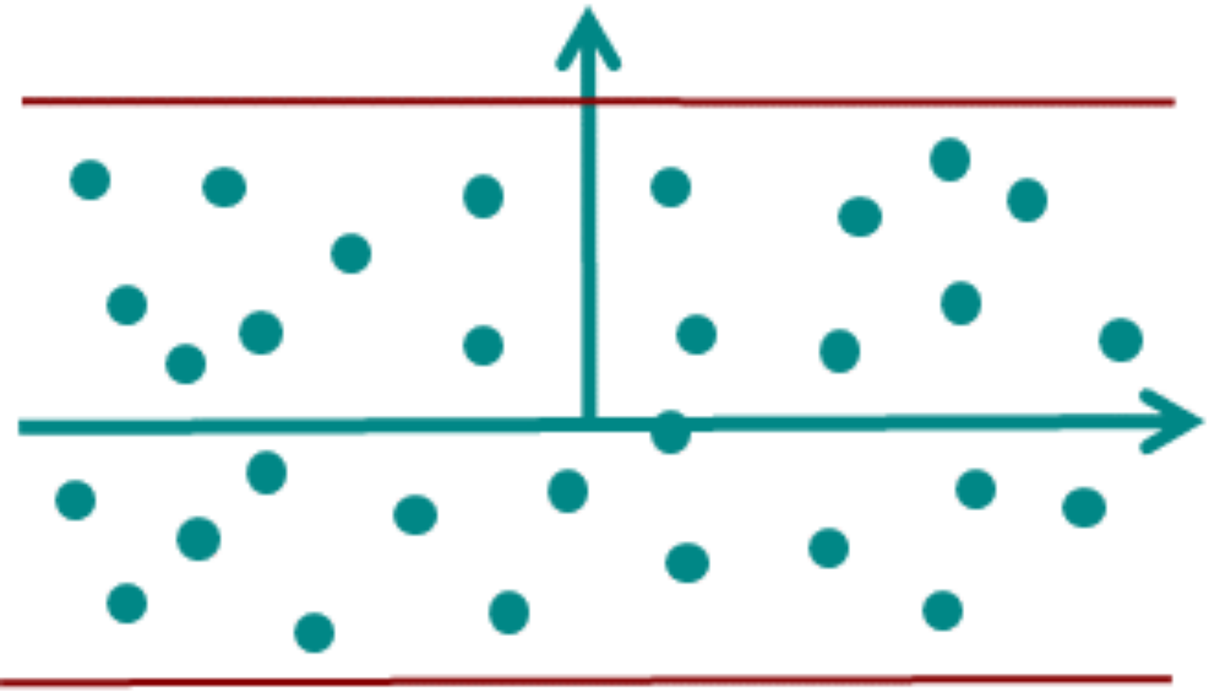
Residuenplots zeigen, ob zentrale Annahmen des linearen Modells verletzt sind.

- Unabhängigkeit: keine Muster entlang der Zeit oder Reihenfolge
- Linearität: Residuen sollten zufällig um 0 streuen
- Homoskedastizität: gleiche Streuung über alle Fits hinweg
- Normalität: relevante Annahme für Inferenz (t-Tests, Konfidenzintervalle)

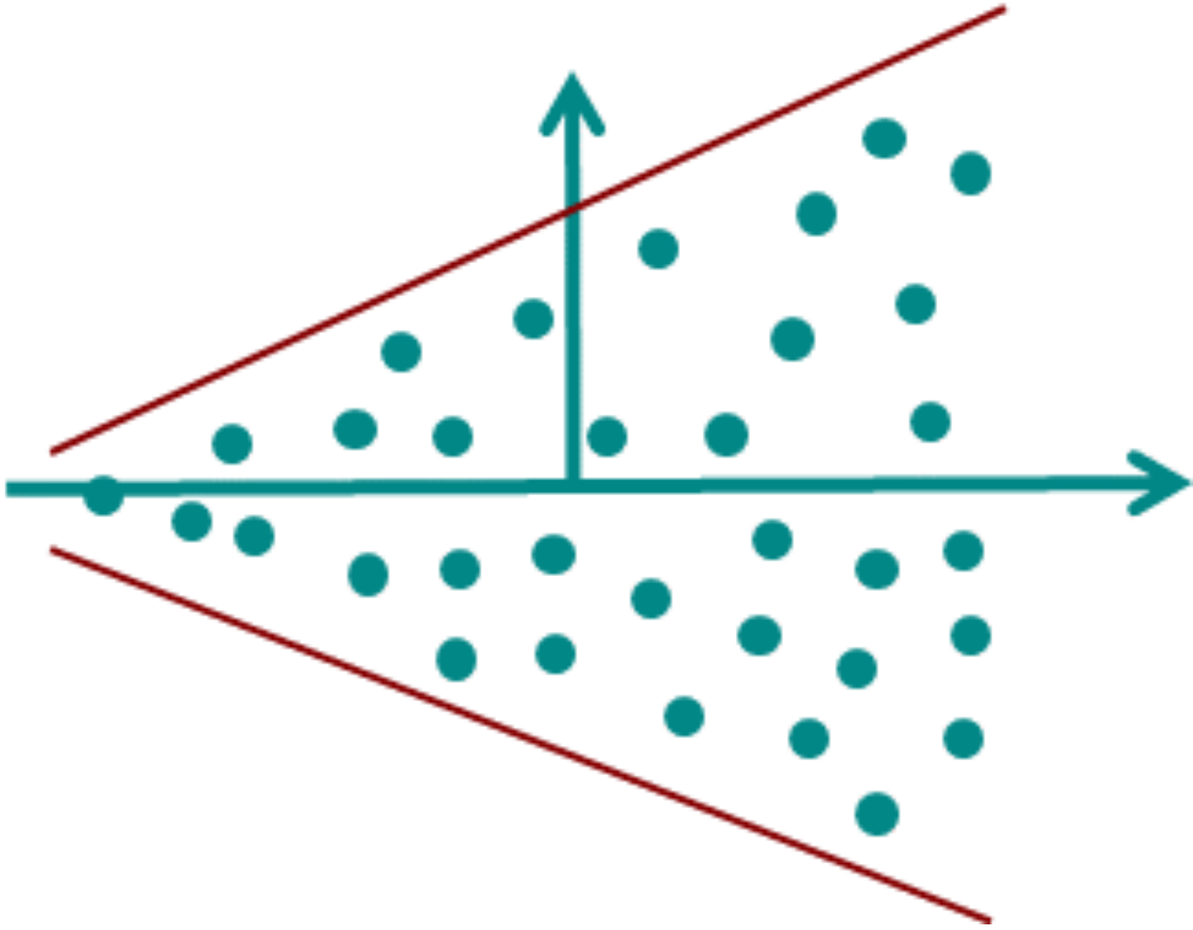
Mini-Check: Welches Muster im Residuen-vs-Fit-Plot deutet auf Heteroskedastizität hin?


Residuals vs Fitted – Heteroskedastizität (Trichterform)
Residuals (eᵢ)
Fitted Values (ŷ)
8
6
4
2
0
−2
−4
−6
2
3
4
5
6
7
8
9

Homoskedastizität



Heteroskedastizität



Zusammenhang Diagnoseplots & Modellgüte

Diagnoseplots zeigen, ob die Annahmen der linearen Regression tragfähig sind.

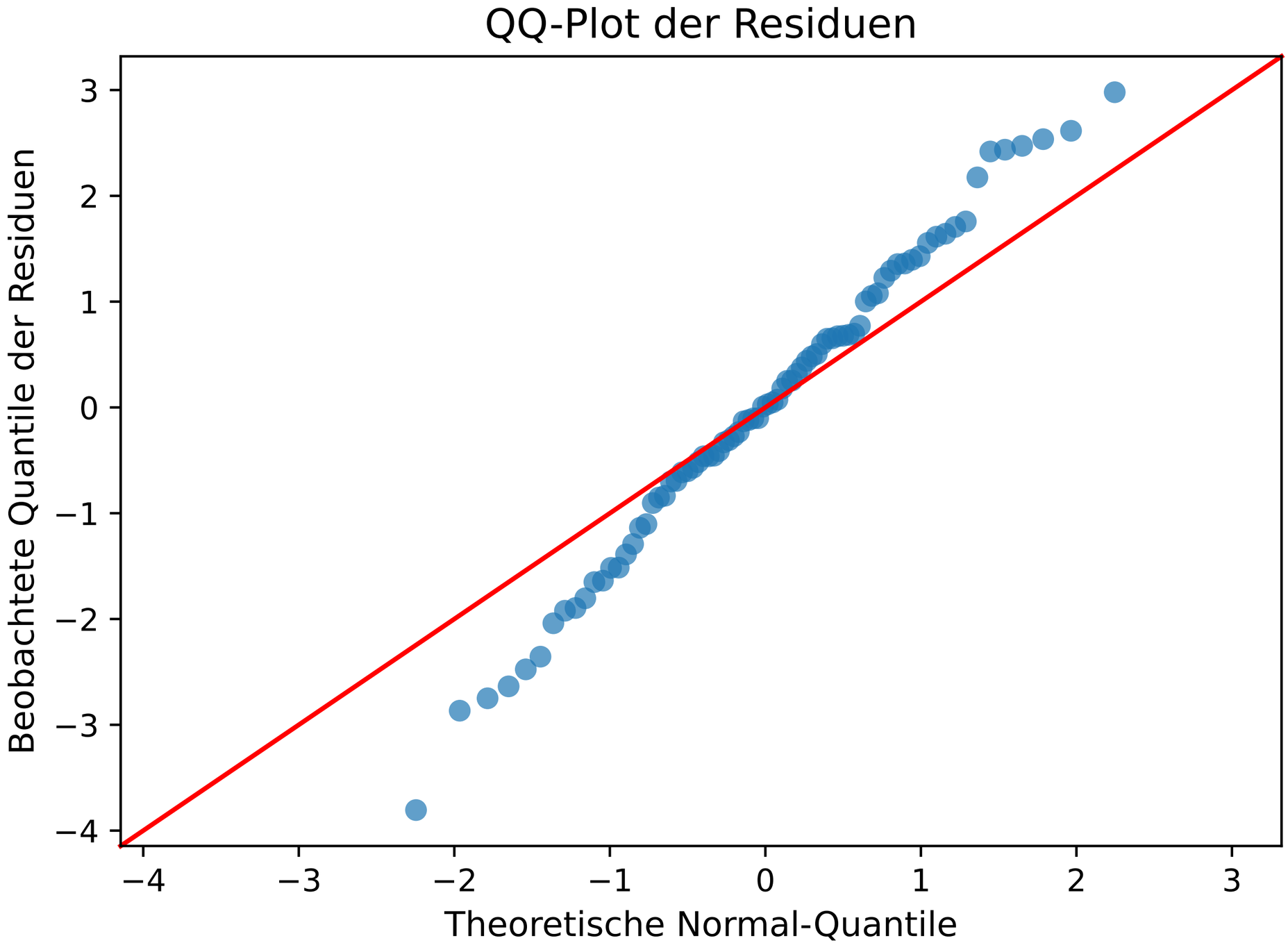
- QQ-Plot: prüft Normalverteilung der Residuen.
- Cook’s Distance: identifiziert einflussreiche Punkte.

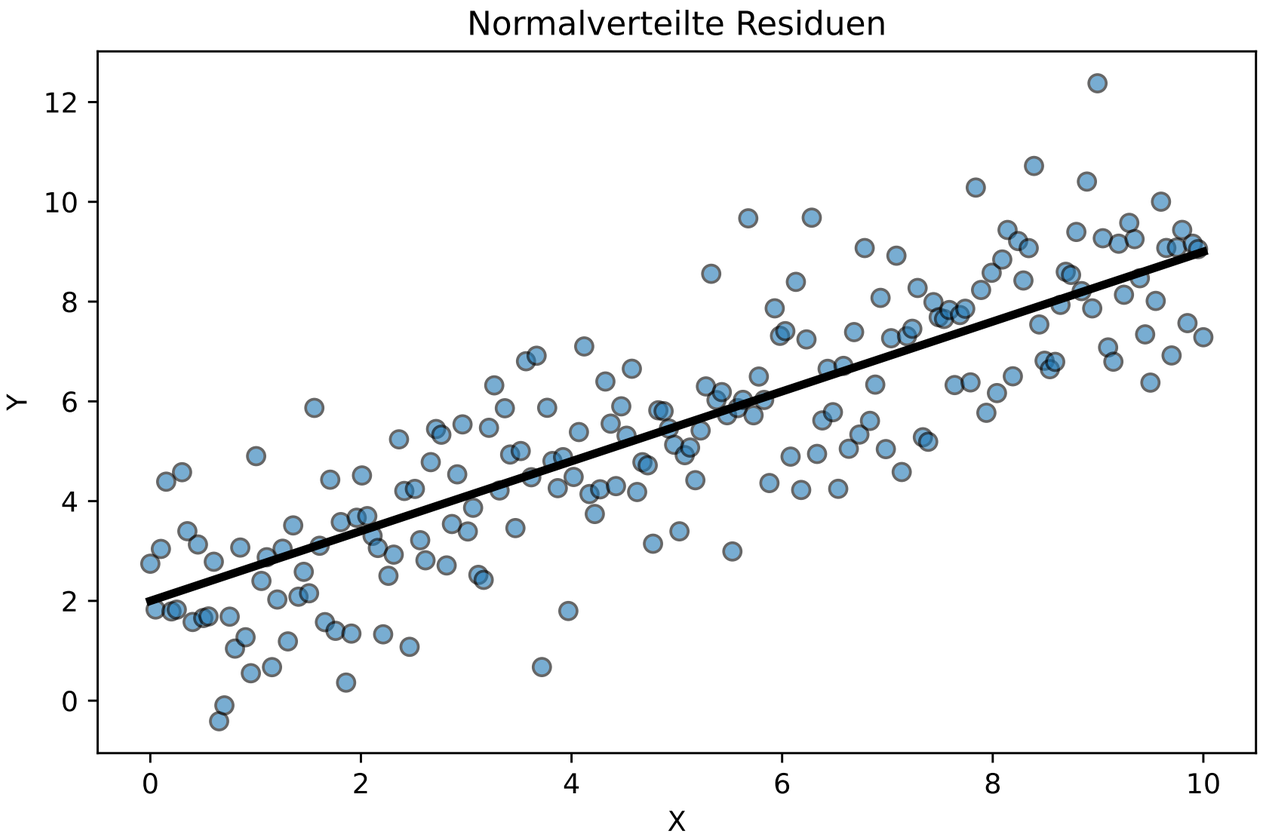
QQ-Plot der Residuen

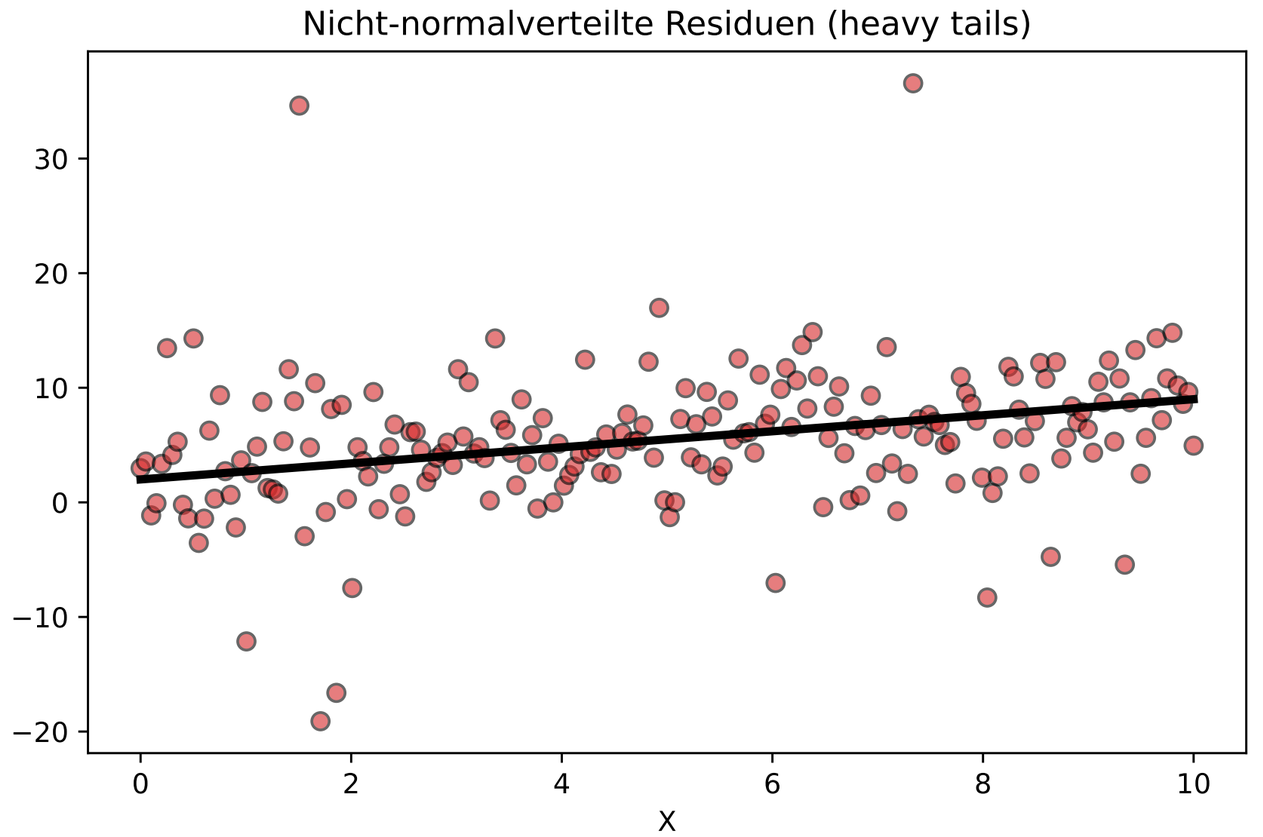
Der QQ-Plot prüft, ob die Residuen annähernd normalverteilt sind.

- X-Achse: theoretische Quantile einer Normalverteilung
- Y-Achse: beobachtete Quantile der Residuen
- Punkte nahe an der Diagonalen → Normalität plausibel
- starke Krümmung oder S-Form → Abweichung von Normalität

Mini-Check: Warum ist Normalität der Residuen wichtig für t-Tests und Konfidenzintervalle?


QQ-Plot der Residuen
Beobachtete Quantile der Residuen
Theoretische Normal-Quantile
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3


Normalverteilte Residuen
Y
X
0
2
4
6
8
10
12

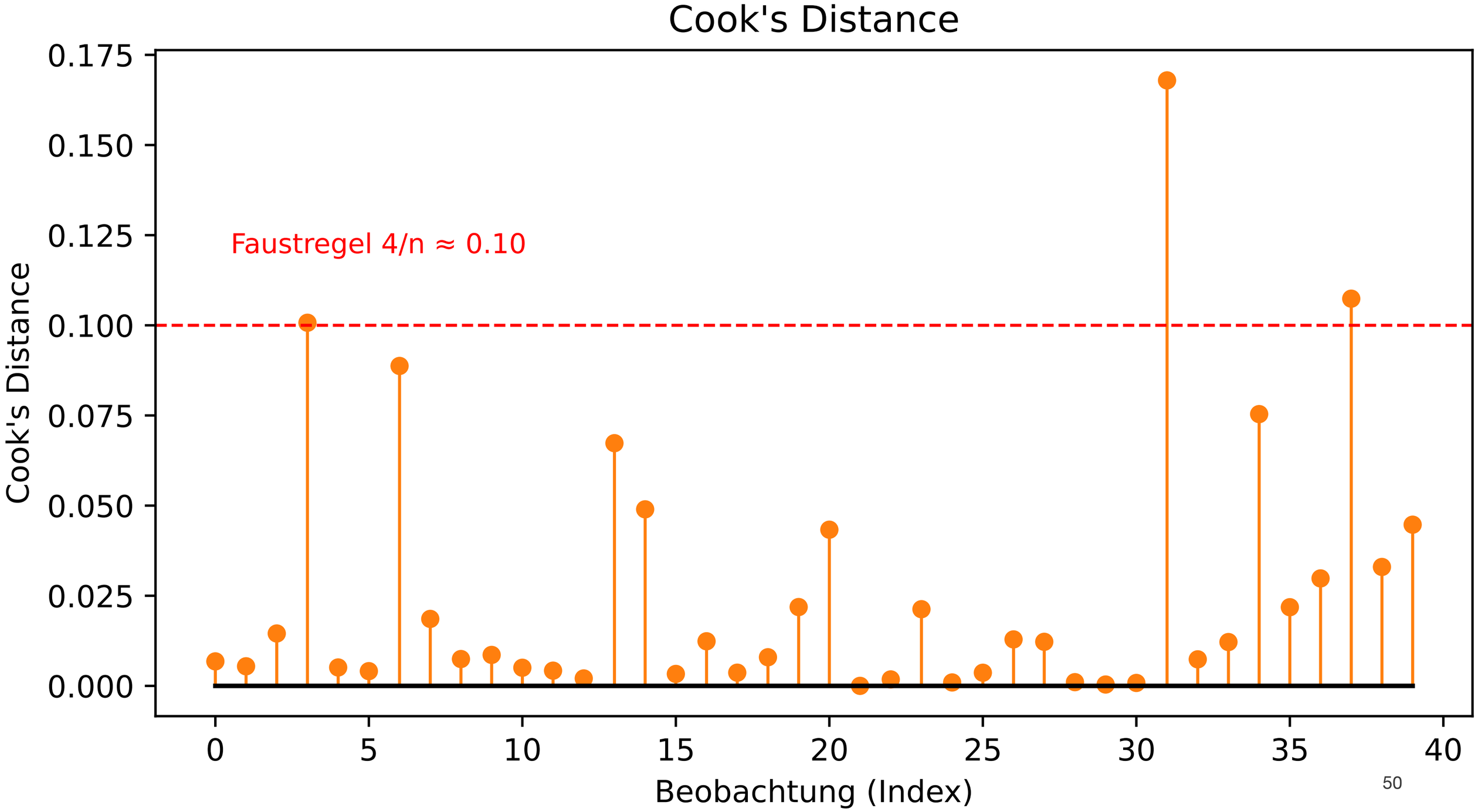

Nicht-normalverteilte Residuen (heavy tails)
X
−20
−10
0
10
20
30

Cook’s Distance

Cook’s Distance identifiziert Beobachtungen mit grossem Einfluss auf das Modell.

- misst kombiniert:
  - Grösse des Residuums
  - Leverage (Lage im X-Raum)
- grosse Cook’s-Werte → Punkte, die die Regressionsgerade stark verändern
- Faustregel: Werte deutlich über dem Rest genauer prüfen

Mini-Check: Warum kann ein Punkt mit kleinem Residuum trotzdem eine grosse Cook’s Distance haben?


Cook's Distance
Faustregel 4/n ≈ 0.10
Cook's Distance
Beobachtung (Index)
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25
30
35
40

Take-Away: Residuenplots & Modellannahmen

Residuenplots sind das zentrale Werkzeug, um Modellprobleme sichtbar zu machen.

- Residuen-vs-Fit: zeigt Muster, Nichtlinearität, Heteroskedastizität
- QQ-Plot: prüft Normalverteilung der Residuen
- Cook’s Distance: identifiziert einflussreiche Punkte (Leverage + Residuum)

Mini-Check: Welche Diagnose prüfst du immer zuerst?

Zusammenfassung

Was Regression nicht kann

- Keine Kausalität beweisen
- Ausreisser automatisch abfangen
- Starke Nichtlinearitäten korrekt modellieren
- Komplexe Zusammenhänge ohne Diagnoseplots sichtbar machen

Mini-Check: Welcher dieser Punkte wird am häufigsten falsch verstanden?

Take-Aways

Regression modelliert, wie sich Y verändert, wenn sich X verändert.

- OLS schätzt die beste Gerade durch Minimierung des RSS.
- R^2 zeigt, wie viel Variation im Y erklärt wird («Güte des Modells»).
- Residuen machen Probleme sichtbar; Diagnoseplots sind Pflicht.
- Einfachheit vor Komplexität: Modelle nicht überladen.
- Interpretation immer im Datenkontext prüfen.

Mini-Check: Was sagt ein R^2 von 0.65 aus?

Was du jetzt kannst

- Regressionsmodell formulieren
- β_0, β_1 interpretieren
- Residuenplots lesen und beurteilen
- RSS, TSS, ESS zerlegen
- R^2 (kritisch!) interpretieren
- Qualität eines linearen Modells einordnen

Mini-Check: Welche Diagnose würdest du als erste prüfen?

Quiz: Aktive Wiederholung

Kahoot Quiz VL10

Ausblick

Nächste Woche: Multiple Regression mehrere Einflussgrössen gleichzeitig

- Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$
- Ziel: Confounding besser verstehen und Effekte sauber trennen
- Themen:
  - Interpretation von Koeffizienten bei mehreren X
  - Multikollinearität und VIF
  - Modellvergleich (AIC/BIC) für komplexere Modelle

Glossar

Glossar: Zentrale Regressionsbegriffe

- Regression: Modell, das beschreibt, wie sich Y verändert, wenn sich X verändert.
- Prädiktor / Kovariate: erklärende Variable X im Regressionsmodell.
- Antwortvariable: Zielgrösse Y, die modelliert / vorhergesagt werden soll.
- Lineares Modell:
  $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 (Achsenabschnitt): erwarteter Wert von Y, wenn $X = 0$.
- β_1 (Steigung): Veränderung von Y pro Einheit X, ceteris paribus.
- Residuum e_i: Differenz zwischen beobachtetem und vorhergesagtem Wert,
  $e_i = y_i - \hat{y}_i$.

Glossar: Gütemasse & Zerlegung

- RSS (Residual Sum of Squares):
  $RSS = \sum (y_i - \hat{y}_i)^2$
  unerklärte Variation; je kleiner, desto besser.
- TSS (Total Sum of Squares):
  $TSS = \sum (y_i - \bar{y})^2$
  gesamte Variation von Y um den Mittelwert.
- ESS (Explained Sum of Squares):
  $ESS = \sum (\hat{y}_i - \bar{y})^2$
  erklärte Variation des Modells.
- Zerlegung: $TSS = ESS + RSS$
- Bestimmtheitsmass R^2:
  $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$
  Anteil der erklärten Variation.
- RSE (Residual Standard Error):
  $RSE = \sqrt{\frac{RSS}{n-2}}$
  typische Abweichung eines Punkts von der Geraden.

Glossar: Modellannahmen

Zentrale Annahmen der einfachen linearen Regression:

- Linearität: Zusammenhang zwischen X und Y ist (annähernd) linear.
- Unabhängigkeit: Fehler / Residuen sind unabhängig (keine Zeit-/Reihenfolgemuster).
- Homoskedastizität: Varianz der Residuen ist für alle $\hat{y}$ etwa gleich.
- Normalität der Fehler: Residuen sind (annähernd) normalverteilt.

Konsequenzen:

- Verletzte Annahmen → verzerrte Standardfehler, Tests, Konfidenzintervalle.
- Diagnoseplots helfen, diese Annahmen zu prüfen.

Glossar: Diagnoseplots

Wichtige Diagnoseplots in der Regression:

Residuen-vs-Fit: Residuen gegen $\hat{y}$ geplottet; zeigt

- Muster (Nichtlinearität)
- Trichterform (Heteroskedastizität)
- Ausreisser (vertikale Ausreisser).

QQ-Plot: Residuen-Quantile vs. Normal-Quantile;

- Punkte auf Diagonale → Normalität plausibel
- S-Form oder starke Krümmung → Abweichung.

Cook's Distance: misst Einfluss einzelner Punkte (Leverage + Residuum).

Glossar: Einfluss & Leverage

Leverage: Mass dafür, wie weit eine Beobachtung im X-Raum vom „Zentrum“ liegt.

- hohe Leverage-Punkte: extreme x_i
- können die Regressionsgerade stark ziehen.

Einflussreicher Punkt:
Beobachtung, die die geschätzte Gerade stark verändert.

- Kombination aus hohem Leverage und grossem Residuum.

Cook’s Distance D_i:
kombiniert Residuum und Leverage;

- grosse Werte → Punkt hat starken Einfluss auf die Schätzung.
- Faustregel: $D_i > 4/n$ genauer prüfen.