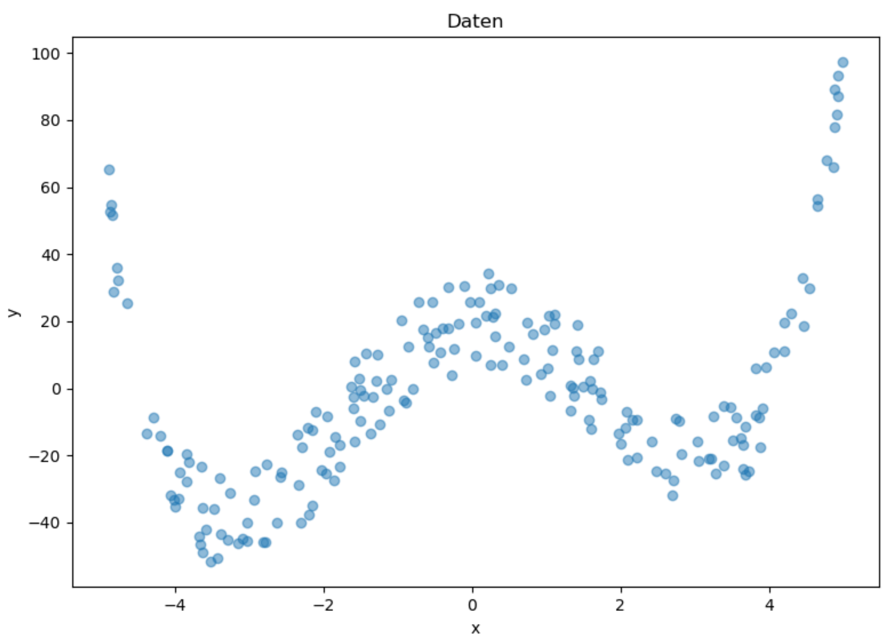
Mathematisch-algorithmische Grundlagen für Data Science

Lösungen zu Aufgabenblatt 5

Aufgabe 1 (Lineare Ausgleichsrechnung) — 2+5+1=8 Punkte

(a) Erstellen Sie einen Scatterplot für die Daten von Aufgabe 1 und beantworten Sie damit die folgende Frage:

> Wenn Sie als Ansatz für eine Ausgleichsfunktion ein Polynom wählen, welchen Grad sollte dieses Polynom mindestens haben?



Da drei Extremstellen vorzuliegen scheinen, muss ein Regressionspolynom mindestens den Grad 4 haben.

(b) Sie entscheiden sich gemäß (a) für ein Polynom mit minimal sinnvollem Grad als Ansatz für eine Ausgleichungsfunktion (Regressionspolynom).

Stellen Sie das Normalgleichungssystem auf und berechnen Sie damit das Regressionspolynom.

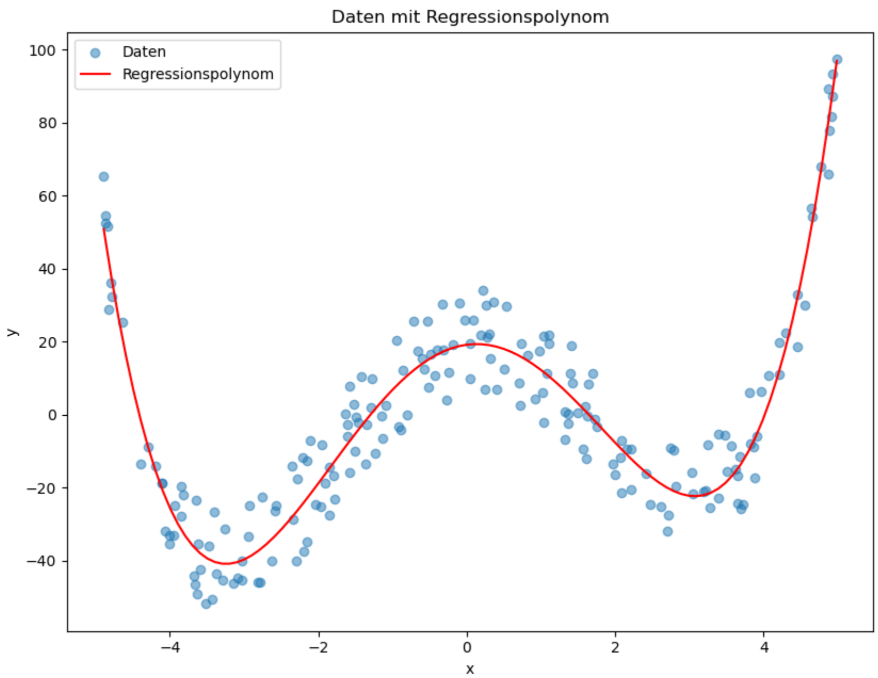
Für ein Polynom

$$p(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

ergeben sich folgende Koeffizientenwerte:

$$\begin{aligned} a_4 &= 0.50702238 \\ a_3 &= 0.01965818 \\ a_2 &= -10.13358585 \\ a_1 &= 2.73895381 \\ a_0 &= 19.15283518 \end{aligned}$$

(c) Erweitern Sie den Plot aus (a) um das berechnete Regressionspolynom.



Aufgabe 2 (Polynominterpolation vs. lineare Regression) 6 Punkte

Bei einer zufällig ausgewählten Gruppe von Zuschauern an einem Basketballspiel kamen die nebenstehenden Messungen zustande.

Körpergröße	Gewicht
1.55	51
1.57	50
1.62	55
1.68	52
1.75	60
1.76	68
1.81	78
1.83	91
1.87	84
1.89	81
1.90	90
1.92	105
1.95	95
1.96	99
1.99	100
2.02	101

Schreiben Sie ein Programm, das folgendes leistet:

- Es wählt von den 16 Datensätzen acht Datensätze zufällig aus,
- berechnet zu diesen acht Datensätzen die Regressionsgerade,

- berechnet weiterhin ein Polynom siebten Grades, das exakt durch die ausgewählten acht Datenpunkte verläuft und
- ermittelt sowohl für die Regressionsgerade als auch für das Polynom den RMSE (Root Mean Squared Error) für die *nicht* ausgewählten acht Datensätze.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}$$

Hierbei ist f die Ausgleichsfunktion, also die Regressionsgerade bzw. das Interpolationspolynom.

Lassen Sie Ihr Programm mehrmals laufen und schauen Sie sich die Ergebnisse an.

Welche Methode ist für die Prognose des Gewichts auf Basis der Körpergröße besser geeignet?

Ergebnisse (gerundet) von fünf Durchläufen:

Regressionsgerade	Interpolationspolynom
8.34	276.42
8.99	49.52
6.61	106.12
5.94	34.44
7.77	85.34

Bei weiteren Durchläufen ergeben sich teilweise extrem große RMSE-Werte für das Interpolationspolynom. Demgegenüber sind die RMSE-Werte der Regressionsgerade stabil und deutlich kleiner.

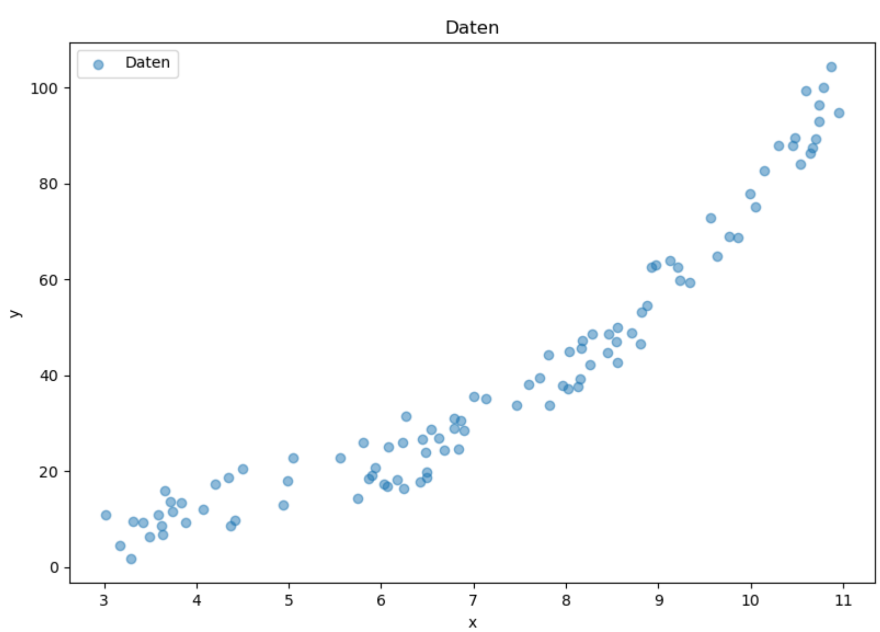
Die Regressionsgerade als Methode ist deutlich besser, die Methode mit dem Interpolationspolynom ist unbrauchbar.

Aufgabe 3 (Logarithmisch-lineare Regression) 1+5+1=7 Punkte

Für die gegebenen Daten vermuten wir folgenden funktionalen Zusammenhang:

$$f(x) = \alpha\, x^{\beta}\, \mathrm{e}^{\gamma x}.$$

(a) Erstellen Sie einen Plot der Daten.



(b) Ermitteln Sie eine Ausgleichsfunktion mittels logarithmisch-linearer Regression.

Es ergibt sich

$$\begin{aligned} \alpha &= 1.705725028697076 \\ \beta &= 0.6421408025086609 \\ \gamma &= 0.23090791744913278 \end{aligned}$$

und damit (gerundet)

$$f(x) = 1.71\, x^{0.64}\, \mathrm{e}^{0.23x}.$$

(c) Erweitern Sie Ihren Plot aus (a) um die Ausgleichsfunktion.