

Methode der kleinsten Quadrate

Unterlagen für Lehrende

Miriam Schönauer, MSc

23. Juni 2023

Grundsätzlich: Die Methode der kleinsten Quadrate kann verwendet werden, um das Verhältnis zweier Variablen zu modellieren und damit Vorhersagen zu treffen.

Technologieeinsatz mit Python:

Link zu Replit: 1. <https://replit.com/join/odwxxskclt-miriamscho>

2. <https://replit.com/join/rgrbafad-miriamscho>

3. <https://replit.com/join/vwwpxgfigo-miriamscho>

Einige nützliche Python-Befehle: numpy library: `numpy.sum(x)`, `numpy.size`

Programm in Stichworten:

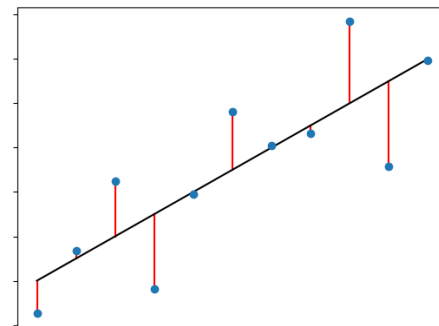
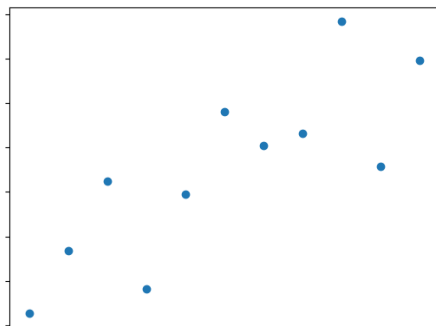
Wiederholung Ableitung; Methode der kleinsten Quadrate (Linear) : Problemstellung und Herleitung, Anwendungsbereich, Implementierung in Python, Experimente mit verschiedenen Datensätzen, Motivation polynomieller Least Squares Methoden, Herleitung, Anwendung auf verschiedene Probleme

Dauer: 2 Unterrichtseinheiten; Voraussetzung: Grundlegende Python-Kenntnisse, Lösen von Gleichungssystemen, Kenntnis der Ableitung; Niveau: ★★

Es seien Werte $y_1, \dots, y_n \in \mathbb{R}$ (z.B. Messungen oder Funktionsauswertungen) zu endlich vielen Punkten $x_1, \dots, x_n \in D$ vorgegeben. Es soll nun eine Funktion $\ell : D \rightarrow \mathbb{R}$ aus einer Klasse L von Funktionen berechnet werden, die diese Daten möglichst gut approximiert. Dabei muss nicht zwingend gelten, dass $\ell(x_i) = y_i$. Es soll vielmehr das Verhältnis zwischen den x_i -Werten und den y_i -Werten modelliert werden. Als Klasse L verwendet man sehr häufig die linearen Funktionen, da diese Funktionen sehr einfach zu handhaben sind. Es bezeichne

$$L := \{\ell(x) = a_0 + a_1x : a_i \in \mathbb{R}\}$$

den Raum der linearen Funktionen. Es soll also ein $\ell \in L$ ermittelt werden, sodass ℓ möglichst "gut" zu den Daten passt. Dafür muss festgelegt werden, was als "gute" Approximation zu verstehen ist. Da bei der Approximation eine Gerade durch Datenpunkte gelegt wird, liegt es nahe, dass eine gute Annäherung an die Daten, eine lineare Funktion bezeichnet, die den senkrechten Abstand zwischen Datenpunkten und Gerade minimiert.



Der Abstand zwischen Gerade und einem Datenpunkt lässt sich formal beschreiben über $|y_i - (a_0 + a_1 x_i)|$. Da wir für alle Punkte einen möglichst kleinen Abstand erreichen möchten, minimieren wir über die Summe der Abstände, das heißt

$$\min_{(a_0, a_1) \in \mathbb{R}^2} \sum_{i=1}^n |y_i - (a_0 + a_1 x_i)|.$$

Dabei minimieren wir über (a_0, a_1) , da sie die Unbekannten in dieser Aufgabe darstellen. Um das Minimum, und somit die Unbekannten (a_0, a_1) zu ermitteln, möchte man nun diesen Ausdruck jeweils in a_0 bzw. a_1 ableiten, um dann durch Nullsetzen dieser Ableitungen ein Gleichungssystem für a_0 und a_1 zu erhalten.

Allerdings gibt es hier ein Problem. Da in dem Summenausdruck die Betragsfunktion vorkommt, kann nicht die herkömmliche Ableitung benutzt werden. Daher wird ein anderer Ausdruck minimiert nämlich

$$\min_{(a_0, a_1) \in \mathbb{R}^2} r(a_0, a_1) = \sum_{i=1}^n (y_i - (a_0 + a_1 x_i))^2.$$

Diese Summe, abgekürzt als Funktion $r(a_0, a_1)$, kann problemlos abgeleitet werden und minimiert die Summe der Quadrate der Abstände. Dieses Minimum ist auch im Sinne der Problemstellung, ändert aber natürlich die Definition von „gut“ bei der Approximation etwas ab.

Die Lösung der Problemstellung, also die „beste“ Gerade für die Datenpunkte, ist durch das Minimum der Fehlerquadrate gegeben. Dieses Prinzip der Minimierung der Quadrate wird auch als Methode der kleinsten Quadrate bezeichnet oder auch als Least Squares Methode.

Jetzt ist es möglich, die Ableitungen von r zu bilden.

$$\begin{aligned} \frac{\partial r}{\partial a_0} &= 2 \sum_{i=1}^n (y_i - (a_0 + a_1 x_i)) \cdot (-1) = 2(a_0 \sum_{i=1}^n 1 + a_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i) \\ \frac{\partial r}{\partial a_1} &= 2 \sum_{i=1}^n (y_i - (a_0 + a_1 x_i)) \cdot (-x_i) = 2(a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i x_i) \end{aligned}$$

Wenn diese Ableitungen gleich 0 gesetzt werden, ergibt sich das folgende Gleichungssystem

$$\begin{aligned} a_0 n + a_1 S_x - S_y &= 0 \\ a_0 S_x + a_1 S_{xx} - S_{xy} &= 0, \end{aligned}$$

wobei $S_x := \sum_{i=1}^n x_i$, $S_y := \sum_{i=1}^n y_i$, $S_{xx} := \sum_{i=1}^n x_i^2$ und $S_{xy} := \sum_{i=1}^n y_i x_i$. Durch einfaches Lösen des Gleichungssystems ergeben sich für die Koeffizienten die Formeln

$$\begin{aligned} a_0 &= \frac{S_y S_{xx} - S_{xy} S_x}{n S_{xx} - S_x^2}, \\ a_1 &= \frac{n S_{xy} - S_x S_y}{n S_{xx} - S_x^2}. \end{aligned}$$

Mit diesen Formeln lassen sich die [Arbeitsaufträge 1-3](#) per Hand lösen. Ziel ist es dabei, sich mit den Formeln und der Vorgehensweise vertraut zu machen.

In [Arbeitsauftrag 4](#) soll jetzt die Implementation auf dem Computer folgen, bevor dann damit die [Arbeitsaufträge 5-11](#) bearbeitet werden. Dabei ist unter dem 1. Replit-Link eine Python-Datei verfügbar, die das Grundgerüst für die Implementation ist. Hier steht die Experimentation mit echten Daten und das Verständnis im Vordergrund. Es soll selbst erarbeitet werden, wo die lineare Least Squares Methode gut funktioniert und wo sie an ihre Grenzen stößt. Auch soll vermittelt werden, dass die graphische Darstellung der Daten und der Lösung sehr wichtig sind in der Analyse des Modells (auch bei komplexeren numerischen Verfahren ist die Visualisierung der Modellierung ein sehr wichtiges Werkzeug).

Falls es bei der eigenständigen Implementation zu Schwierigkeiten kommt, ist unter dem 2. Replit-Link eine Python-File verfügbar, in der bereits die Methode `koeff_bestimmen(x, y)` implementiert ist.

Bei den [Arbeitsaufträgen 7-10](#) fällt auf, dass die Least Squares Gerade die Daten nur zum Teil bzw. gar nicht approximiert. Die Zusammenhänge zwischen den Daten ist nicht immer linear, wie bei diesen Beispielen hervorgeht. Also ist die Klasse L nicht hinreichend für diese Zwecke.

Um dieses Problem zu lösen, können wir dazu übergehen, eine etwas komplexere Klasse an Funktionen für die Modellierung zu verwenden. Die Polynomfunktionen bieten sich an, da sie zwar „einfache“ Funktionen sind, aber dennoch sehr flexibel sind. Es sei also

$$P_m := \{p(x) = a_0 + a_1x + \dots + a_mx^m : a_i \in \mathbb{R}, i = 0, \dots, m\}$$

der Raum der Polynome vom Grad kleiner oder gleich m .

Es wird noch immer die Summe der Quadrate der Abstände minimiert. Es ist also gesucht

$$\min_{(a_0, a_1) \in \mathbb{R}^2} r(a_0, \dots, a_m) = \sum_{i=1}^n (y_i - (a_0 + \dots + a_mx_i^m))^2.$$

Um die nötigen Gleichungen zur Bestimmung der Koeffizienten $a_0, \dots, a_m$ zu erhalten, muss die Summe nun für jeden Koeffizienten abgeleitet und gleich 0 gesetzt werden. (Hier kann zur Veranschaulichung das Gleichungssystem für $p(x) = a_0 + a_1x + a_2x^2$ hergeleitet werden.)

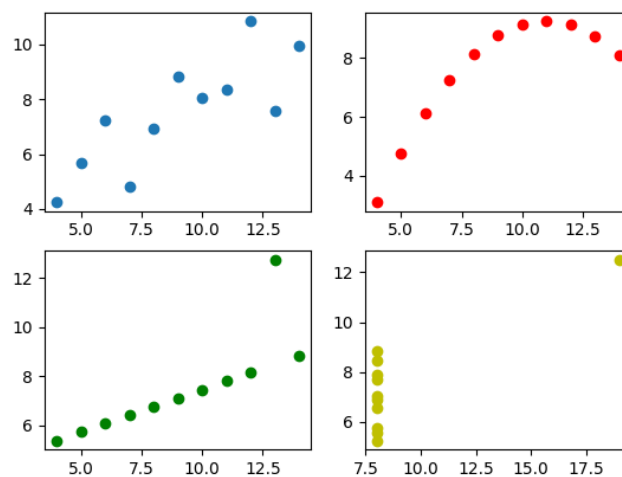
Für die [Arbeitsaufträge 11-13](#), die nun mit polynomieller Least Squares Methode gelöst werden sollen, soll die bereits implementierte Funktion `LS_poly` verwendet werden. Ziel ist es bei diesen Aufgaben zu sehen, wie gut nun der polynomielle Ansatz funktioniert und wie sehr das Ergebnis vom gewählten Grad der Funktion abhängt. Es soll auch experimentiert werden: Was passiert, wenn der Grad des Polynoms sehr hoch gewählt wird?

Für alle Aufgaben ist unter dem 3. Replit-Link ein Lösungsvorschlag zur Verfügung gestellt.

Arbeitsaufträge:

1. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 1), (2, 3)$ und $(3, 3)$ per Hand.
2. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 0), (4, 2), (5, 1)$ und $(7, 2)$ per Hand.
3. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 0), (0, 2)$ und $(2, 2)$. Vertausche dann die Rollen von x und y und berechne die Least Squares Gerade. Skizziere beide Least Squares Lösungen und überlege, welcher Abstand minimiert wird.
4. Implementiere die Berechnung der Least Squares Koeffizienten in der Funktion `koeff_bestimmen(x, y)` in Python. Kontrolliere den Code mit den Ergebnissen aus den Aufgaben 1 und 2.

5. Lade die *marketing.csv* Datei. In dieser Datei sind die Ausgaben für TV-Werbungen (in Millionen), Social Media Werbungen und Radio Werbungen den Verkaufszahlen eines Produkts (in Millionen) gegenübergestellt. Modelliere das Verhältnis zwischen den Ausgaben und den Verkaufszahlen. Schätze jeweils die Verkaufszahlen für ein Budget von 50 Millionen. Für welche Werbungen modelliert die Least Squares Lösung das Verhalten am besten?
6. Lade die Datei *Salary_Data.csv*. In dieser Datei sind die Berufserfahrung (in Jahren) und das Jahresgehalt gegenübergestellt. Modelliere das Verhältnis zwischen der Berufserfahrung und dem Jahresgehalt. Schätze wie hoch das Jahresgehalt ist für 12 und 15 Jahre.
7. Lade die *abalone.csv* Datei. In dieser Datei sind mehrere Körpermaße von Abalonen ihrem Alter gegenübergestellt. Modelliere das Verhältnis zwischen den einzelnen Körpergrößen und des Alters. Für welchen Parameter modelliert die Least Squares Lösung das Verhältnis am besten? Modelliere noch weitere Verhältnisse wie zum Beispiel zwischen der Länge und dem Gewicht der Muschelschale („shell_wt“). Fällt dir dabei etwas auf?
8. Berechne die Least Squares Gerade der Daten $(x, y) = (-1, 4), (-0.5, 3), (0, 2), (1, 3), (2, 5)$ und $(3, 10)$.
9. Berechne die Least Squares Gerade der Daten $(x, y) = (1, 100), (3, 80), (6, 60), (8, 60), (10, 70), (13, 75), (15, 78), (18, 90)$ und $(21, 99)$.
10. **Anscombe-Quartett**²: Gegeben sind vier Datensätze:



Außerdem sei $g(x) = 0,5 \cdot x + 3$ eine Gerade. Vermute, welcher Datensatz diese Gerade als Least Squares Lösung liefert. Überprüfe dann deine Vermutung mithilfe des Codes. Lade dafür die Datei *AnscombeQuartett.csv* mit `df = pd.read_csv('AnscombeQuartett.csv', sep=';', decimal=',')`. Was fällt dir auf?

11. Berechne die Least Squares Lösung von Aufgabe (8) mit *LS.poly*. Welcher Grad liefert das beste Ergebnis? Was passiert wenn man den Grad erhöht?

12. Berechne die Least Squares Lösung von Aufgabe (9) mit LS_{poly} . Welcher Grad liefert das beste Ergebnis? Was passiert wenn man den Grad erhöht?
13. Lade die *Crab.csv* Datei. Modelliere das Verhältnis zwischen „Length“ und „Weight“, „Weight“ und „Shell Weight“, „Weight“ und „Height“, „Diameter“ und „Weight“. Welche Modellierung liefert das beste Ergebnis?

Literatur:

1. Rannacher, Rolf: Numerik 0: Einführung in die Numerische Mathematik, Heidelberg: Heidelberg University Publishing, 2017. <https://doi.org/10.17885/heiup.206.281>
2. F. J. Anscombe (1973) Graphs in Statistical Analysis, The American Statistician, 27:1, 17-21, DOI: 10.1080/00031305.1973.10478966
3. Public Domain Datensätze bezogen über <https://www.kaggle.com/>
marketing.csv - <https://www.kaggle.com/datasets/harrimansaragih/dummy-advertising-and-sales-data>
Salary_Data.csv - <https://www.kaggle.com/datasets/krishnaraj30/salary-prediction-data-simple-linear-regression>
abalone.csv - <https://www.kaggle.com/datasets/devzohaib/estimating-the-age-of-abalone-at-a-seafood-farm>
Crab.csv - <https://www.kaggle.com/datasets/sidhus/crab-age-prediction>

Methode der kleinsten Quadrate

Handout

Miriam Schönauer, MSc

23. Juni 2023

Grundsätzlich: Die Methode der kleinsten Quadrate kann verwendet werden, um das Verhältnis zweier Variablen zu modellieren und damit Vorhersagen zu treffen.

Technologieeinsatz mit Python:

Link zu Replit: 1. <https://replit.com/join/odwxxskclt-miriamscho>

2. <https://replit.com/join/rgrbbafad-miriamscho>

Einige nützliche Python-Befehle: numpy library: `numpy.sum(x)`, `numpy.size`

Programm in Stichworten:

Wiederholung Ableitung; Methode der kleinsten Quadrate (Linear) : Problemstellung und Herleitung, Anwendungsbereich, Implementierung in Python, Experimente mit verschiedenen Datensätzen, Motivation polynomieller Least Squares Methoden, Herleitung, Anwendung auf verschiedene Probleme

Arbeitsaufträge:

1. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 1), (2, 3)$ und $(3, 3)$ per Hand.
2. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 0), (4, 2), (5, 1)$ und $(7, 2)$ per Hand.
3. Berechne die Least Squares Gerade der Daten $(x, y) = (0, 0), (0, 2)$ und $(2, 2)$. Vertausche dann die Rollen von x und y und berechne die Least Squares Gerade. Skizziere beide Least Squares Lösungen und überlege, welcher Abstand minimiert wird.
4. (*Optional*) Implementiere die Berechnung der Least Squares Koeffizienten in der Funktion `koeff_bestimmen(x, y)` in Python. Kontrolliere den Code mit den Ergebnissen aus den Aufgaben 1 und 2.

Für die weiteren Arbeitsaufträge müssen Daten aus csv-Dateien geladen werden. Dafür werden Befehle aus der *pandas* library verwendet.

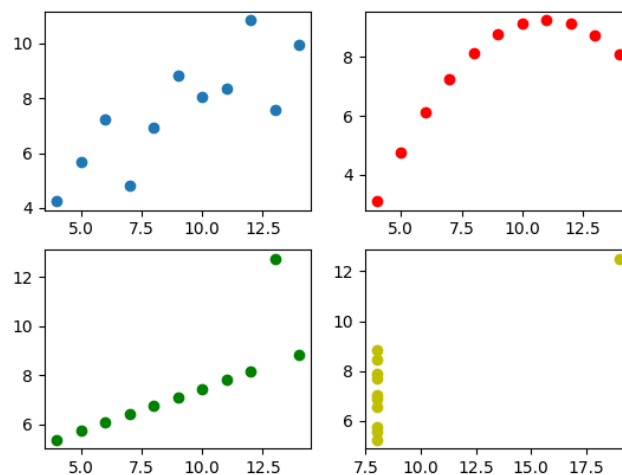
Laden einer Datei : `df = pd.read_csv('marketing.csv', sep=',', header = 0)`

Informationen über den Datei Inhalt: `df.info()`

Dieser Befehl gibt unter anderem die Namen der einzelnen Spalten an, die so Variablen zugeordnet werden können z.B.: $x = df[:, 'TV']$, $y = df[:, 'Sales']$.

5. Lade die *marketing.csv* Datei. In dieser Datei sind die Ausgaben für TV-Werbungen (in Millionen), Social Media Werbungen und Radio Werbungen den Verkaufszahlen eines Produkts (in Millionen) gegenübergestellt. Modelliere das Verhältnis zwischen den Ausgaben und den Verkaufszahlen. Schätze jeweils die Verkaufszahlen für ein Budget von 50 Millionen. Für welche Werbungen modelliert die Least Squares Lösung das Verhalten am besten?
6. Lade die Datei *Salary_Data.csv*. In dieser Datei sind die Berufserfahrung (in Jahren) und das Jahresgehalt gegenübergestellt. Modelliere das Verhältnis zwischen der Berufserfahrung und dem Jahresgehalt. Schätze wie hoch das Jahresgehalt ist für 12 und 15 Jahre.

7. Lade die *abalone.csv* Datei. In dieser Datei sind mehrere Körpermaße von Abalonen ihrem Alter gegenübergestellt. Modellierte das Verhältnis zwischen den einzelnen Körpergrößen und des Alters. Für welchen Parameter modelliert die Least Squares Lösung das Verhältnis am besten? Modelliere noch weitere Verhältnisse wie zum Beispiel zwischen der Länge und dem Gewicht der Muschelschale („shell_wt“). Fällt dir dabei etwas auf?
8. Berechne die Least Squares Gerade der Daten $(x, y) = (-1, 4), (-0.5, 3), (0, 2), (1, 3), (2, 5)$ und $(3, 10)$.
9. Berechne die Least Squares Gerade der Daten $(x, y) = (1, 100), (3, 80), (6, 60), (8, 60), (10, 70), (13, 75), (15, 78), (18, 90)$ und $(21, 99)$.
10. **Anscombe-Quartett:** Gegeben ist sind die vier Datensätze:



Außerdem sei $g(x) = 0,5 \cdot x + 3$ eine Gerade. Vermute, welcher Datensatz diese Gerade als Least Squares Lösung liefert. Überprüfe dann deine Vermutung mithilfe des Codes. Lade dafür die Datei *AnscombeQuartett.csv* mit

```
df = pd.read_csv('AnscombeQuartett.csv', sep=';', decimal=',')
```

Was fällt dir auf?

11. Berechne die Least Squares Lösung von Aufgabe (8) mit *LS_poly*. Welcher Grad liefert das beste Ergebnis? Was passiert, wenn man den Grad erhöht?
12. Berechne die Least Squares Lösung von Aufgabe (9) mit *LS_poly*. Welcher Grad liefert das beste Ergebnis? Was passiert, wenn man den Grad erhöht?
13. Lade die *Crab.csv* Datei. Modellierte das Verhältnis zwischen „Length“ und „Weight“, „Weight“ und „Shell Weight“, „Weight“ und „Height“, „Diameter“ und „Weight“. Welche Modellierung

liefert das beste Ergebnis?

Für die polynomielle Modellierung müssen hier die x -Werte und y -Werte sortiert werden z.B.

```
df = df.sort_values(, , Weight'')
x = df[, , Weight'']
y = df.sort_values(, , Weight'')[, , Height'']
```

Zudem muss in der Datei *LSMP.py* die Zeile `poly_features = poly.fit_transform(x.reshape(-1, 1))` durch `poly_features = poly.fit_transform(x.values.reshape(-1, 1))` ersetzt werden.



FEEDBACK

Was ist die Methode der kleinsten Quadrate? Wofür kann sie verwendet werden?

Was hast du heute sonst gelernt?

Das hat mir gefallen. Das hat mir nicht gefallen.