Python – Lektion 12
Data Handling und Visualisierung mit NumPy/Pandas

Disclaimer

Die Vertraulichkeitsklasse dieser Daten ist
”intern-erweitert”.
Sie dürfen die Daten als OST-Angehörige nutzen, aber nicht an Dritte weitergeben oder veröffentlichen.

Rückblick

- Pandas
  - Series Objekt
  - DataFrame Objekt

Rückblick

- Pandas
  - Series Objekt

    Datenarray mit zugehörigem Index (Labelindex und **Positionsindex**)

Series Objekt

Index	Artikelnummer
0	"124-503"
1	"495-958"
2	"595-838"
3	"123-030"
4	"938-439"

Rückblick

- Pandas
  - Series Objekt

    Datenarray mit zugehörigem Index (**Labelindex** und Positionsindex)

Series Objekt

Index	Artikelnummer
`"Hammer"`	`"124-503"`
`"Flachzange"`	`"495-958"`
`"Pinzette"`	`"595-838"`
`"Messer"`	`"123-030"`
`"Klebeband"`	`"938-439"`

Rückblick

- Pandas
  - DataFrame Objekt
    Aneinanderreihung von Series Objekten, die sich denselben Index teilen (Label- und Positionsindex)

DataFrame Objekt

Index	Artikelnummer	Preis
`"Hammer"`	`"124-503"`	10.00
`"Flachzange"`	`"495-958"`	24.00
`"Pinzette"`	`"595-838"`	14.90
`"Messer"`	`"123-030"`	8.90
`"Klebeband"`	`"938-439"`	4.90

Heutiges Thema

Data Analysis mit Pandas & Numpy am Beispiel einer linearen Regressionsanalyse

- Regressionsanalyse
  - Statistisches Analyseverfahren
  - Modelliert Beziehungen zwischen Variablen
- Beispiele für Beziehungen zwischen...
  - ...Werbeausgaben und Einnahmen eines Unternehmens
  - ...Medikamentendosierung und Blutdruck eines Patienten
  - ...Düngemittel/Wassermenge und Ernteerträge
  - ...Trainigsmethode und Leistung eines Sportlers

Regressionsanalyse

Mathematisch ausgedrückt:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

y_i	Output, abhängige Variable, Antwort (Ernteerträge)
$x_{i,p}$	Input, unabhängige Variable, Feature (Düngemittel, Wassermenge)
$\beta_0, \beta_1, \ldots \beta_p$	Regressionskoeffizienten
ε_i	Störterm
p	Anzahl unabhängige Variablen
$i = 1 \ldots n$	Anzahl Beobachtungen

Regressionsanalyse

Voraussetzungen/Annahmen:

- Linearer Zusammenhang zwischen x und y
- Fehlerterme sind normalverteilt
- Unabhängigkeit der Beobachtungen und der Variablen
- Fehlerterme haben für jeden Wert von x eine konstante Varianz

Regressionsanalyse

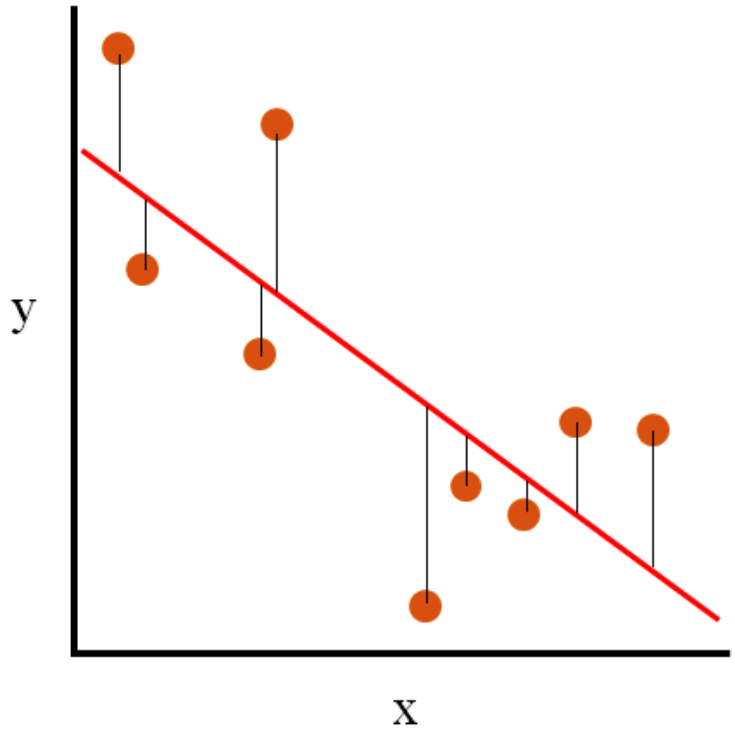
Wir beschränken uns auf eine unabhängige Variable:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Regressionskoeffizienten so berechnen, dass der Fehler ε_i minimal wird:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Grafisch (bei einer unabhängigen Variable):



Regressionsanalyse

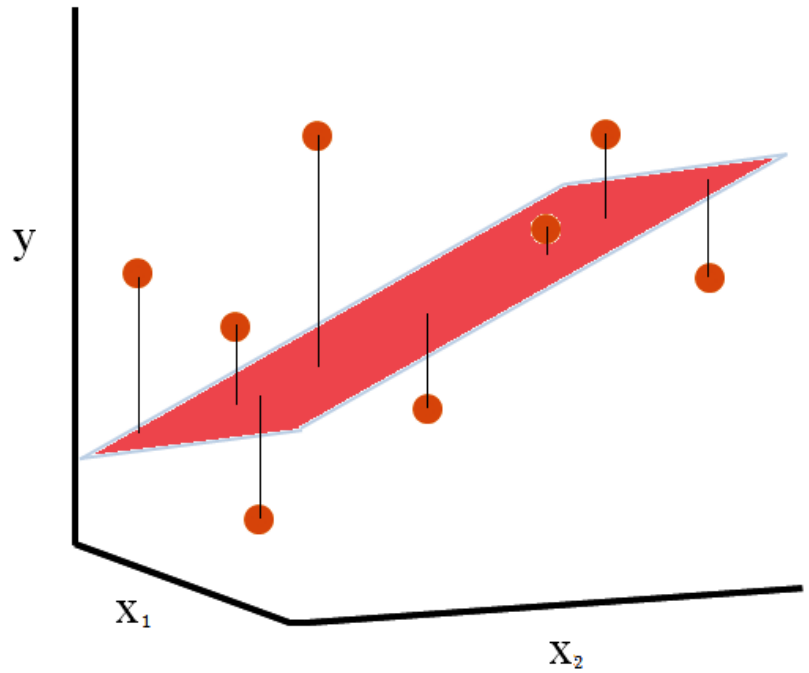
Wir beschränken uns auf eine unabhängige Variable:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Regressionskoeffizienten so berechnen, dass der Fehler ε_i minimal wird:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Grafisch (bei zwei unabhängigen Variablen):



Regressionsanalyse

Wir beschränken uns auf eine unabhängige Variable:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Regressionskoeffizienten so berechnen, dass der Fehler ε_i minimal wird:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Man suche das Minimum der Summe der quadrierten Fehler für alle Beobachtungen $i = 1, \ldots, n$ (OLS - Ordinary Least Squares)

$$\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) \quad \text{fuer} \quad Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Regressionsanalyse

Wir beschränken uns auf eine unabhängige Variable:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Regressionskoeffizienten so berechnen, dass der Fehler ε_i minimal wird:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

Partielle Ableitungen sind beim Minimum Null:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Regressionsanalyse

Lösung für die Regressionskoeffizienten:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

wobei

$$\bar{x}_i = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Regressionsanalyse

`http://localhost:8888/notebooks/regressionsanalyse.ipynb`