

Geodatenanalyse I: Regressionsanalyse – Lineare Regression

Kathrin Menberg



Stundenplan

	08:30 – 12:30 Uhr	13:30 – 17:30 Uhr
Montag	Tag 1 / Block 1	Tag 1 / Block 2
Dienstag	Tag 2 / Block 1	Tag 2 / Block 2
Mittwoch	Tag 3 / Block 1	Tag 3 / Block 2
Donnerstag	Tag 4 / Block 1	Tag 4 / Block 2
Freitag	Tag 5 / Block 1	Tag 5 / Block 2

- ▶ **2.13 Regressionsanalyse – Lineare Regression**
- ▶ 2.14 Regressionsanalyse – Verallgemeinerte lineare Modelle
- ▶ 2.15 Fragestunde und Abschluss

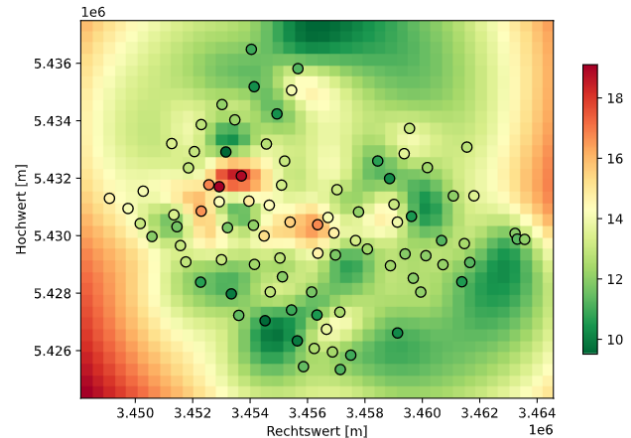
Lernziele Block 2.13

Am Ende der Stunde werden die Teilnehmer:

- ▶ ... mit den mathematischen Grundlagen von der statistischen Regression vertraut sein.
- ▶ ... eine einfache lineare Regression in Python durchführen können.
- ▶ ... die Qualität der Modelanpassung mit Hilfe von verschiedenen Kriterien bestimmen und beurteilen können.

Anknüpfung an gestern...

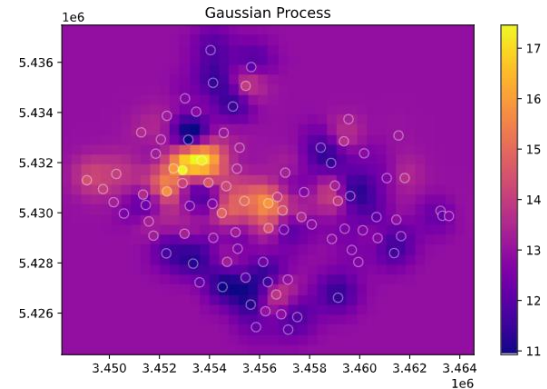
- ▶ Interpolation mit Gauß-Prozess Regression
- ▶ Grundwassertemperatur als Funktion von x- und y-Koordinaten
- ▶ Verallgemeinerung Regression
- ▶ Erklärung einer beobachteten abhängigen Variablen durch eine oder mehrere unabhängige Variablen



Wozu Regressionsanalyse?

- ▶ **Vorhersagen** (prediction)
 - ▶ Modellierung von existierenden Beobachtungen
 - ▶ Neue Datenwerte vorhersagen
 - ▶ Siehe Interpolation mit Gauß-Prozesse zur Regression

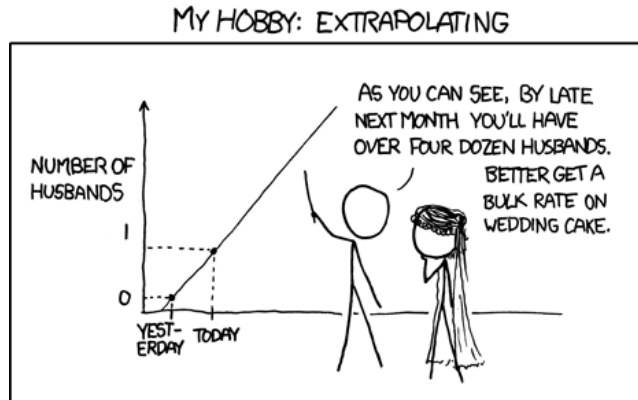
- ▶ **Variablenassoziation**
 - ▶ Zusammenhänge von Variablen identifizieren
 - ▶ Gliederungen und Strukturen in Datensätzen



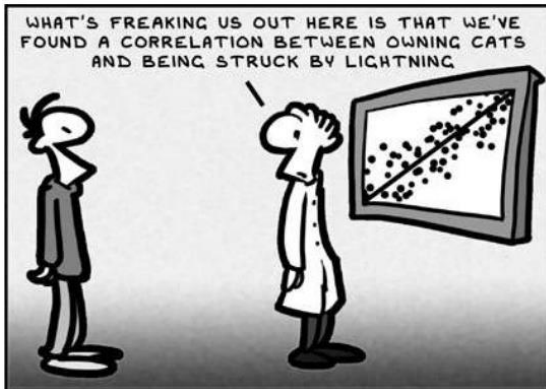
Wozu Regressionsanalyse?

► Extrapolation

- Ausgleichen des Unterschieds zwischen Stichprobe und Grundgesamtheit



www.pinterest.at



► Kausale Schlussfolgerungen

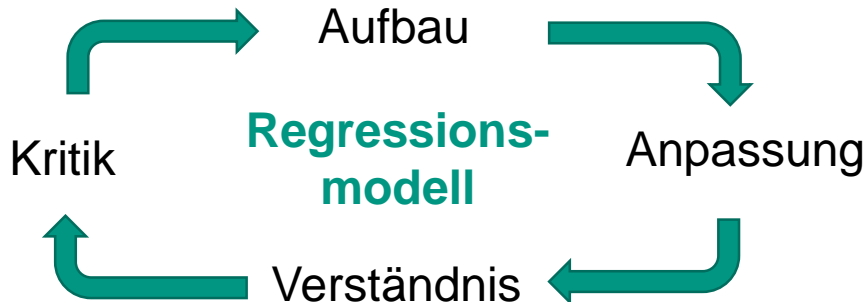
(causal inference)

- Effekte von Verfahren (Variablenänderungen) ableiten
- Experimentelles Design!

4-Stufen Zyklus der statistischen Analyse

- ▶ Schwachstellen suchen
- ▶ Annahmen hinterfragen
- ▶ Mögliche Verbesserungen

- ▶ Modell erweitern
- ▶ Variablen hinzufügen
- ▶ Daten transformieren

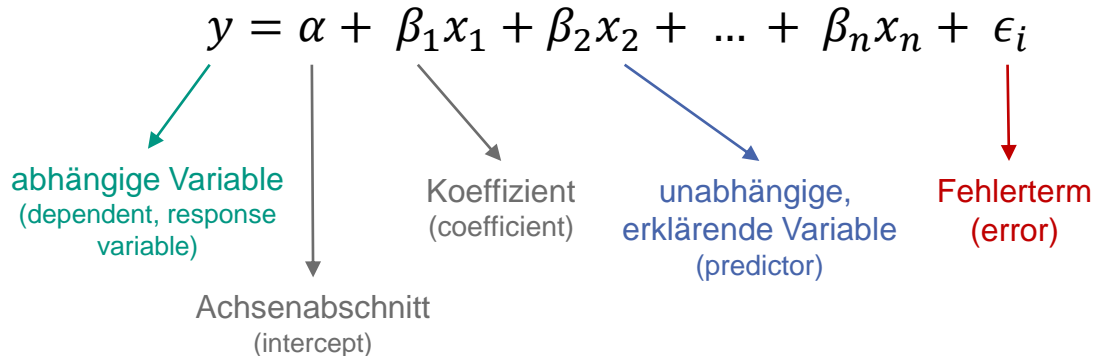


- ▶ Graphische Darstellung
- ▶ Beziehungen zwischen Variablen und Messungen untersuchen

- ▶ Datenmanipulation
- ▶ Koeffizienten schätzen
- ▶ Unsicherheiten

Grundlagen lineare Regression

- ▶ Abhängige Variable als eine Linearkombination der Regressionskoeffizienten

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$$


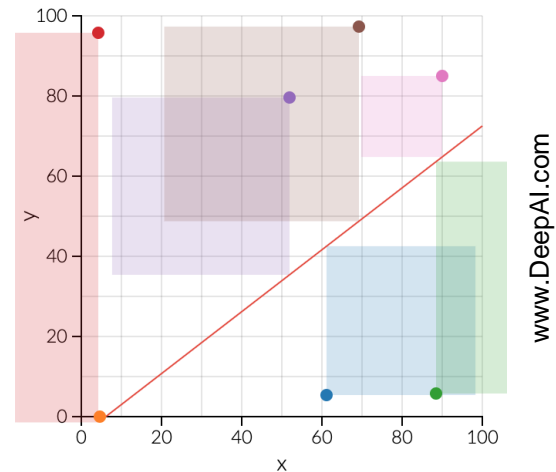
The diagram illustrates the components of the linear regression equation $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$. Arrows point from each term to its description:

- y points to "abhängige Variable (dependent, response variable)" in green.
- α points to "Achsenabschnitt (intercept)" in black.
- β_1 points to "Koeffizient (coefficient)" in black.
- x_1 points to "unabhängige, erklärende Variable (predictor)" in blue.
- ϵ_i points to "Fehlerterm (error)" in red.

- ▶ eine unabhängige Variable: einfache lineare Regression (x_1)
- ▶ mehrere unabhängige Variablen: multiple lineare Regression (x_n)
- ▶ Ziel: Parameter $\hat{\alpha}$ und $\hat{\beta}_i$ finden, die die beste Übereinstimmung zwischen gemessenen und berechneten Werten liefern (ϵ_i minimieren)

Kleinste-Quadrate (KQ) Schätzung

- ▶ engl. Ordinary Least Squares (OLS)
- ▶ Berechnung der Summe der quadrierten Residuen
- ▶ Koeffizienten für einfache lineare Regression:
 - ▶ $\hat{\alpha} = \bar{Y} - \beta * \bar{X}$
 - ▶ $\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{cov(X,Y)}{var(X)}$
- ▶ Für multiple lineare Regression:
 - ▶ $\hat{y} = (X^t X)^{-1} X^t y$
 - ▶ Vektor mit Koeffizienten $\hat{y} = (\alpha, \beta)$



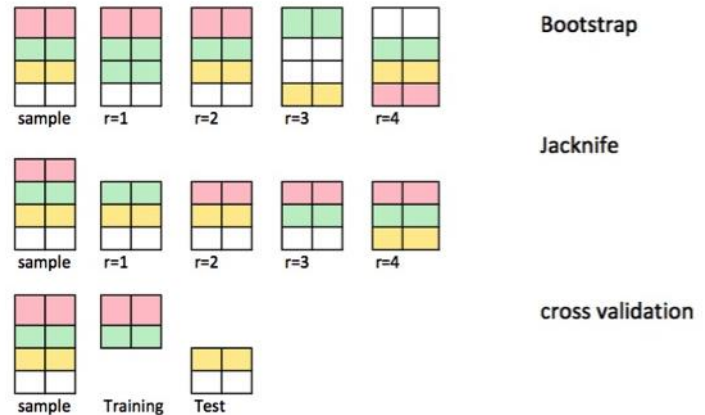
Überprüfung der Anpassungsgüte

► Fehlermaße:

- Root Mean Square Error (RMSE)
- Residuenquadratsumme (SQR) und totale Quadratsumme (SQT)
- Bestimmtheitsmaß (R^2)
- u.v.m.

► Methoden zur Validierung:

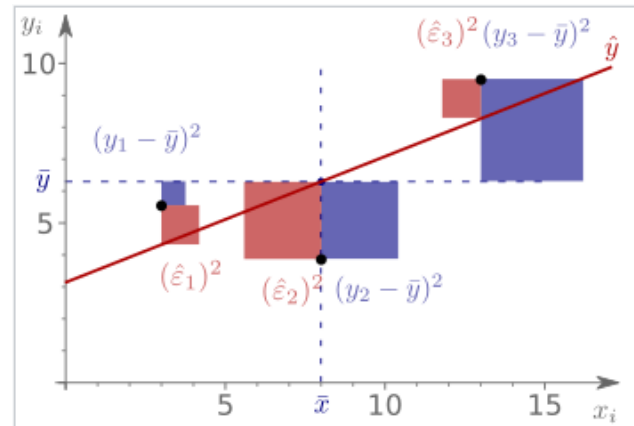
- Bootstrap und Jackknife
- Kreuzvalidierung
- u.v.m.



Fehlermaße

- ▶ y : Beobachtungen, \hat{y}_i Vorhersagen, \bar{y} : Mittelwert der Beobachtungen
- ▶ totale Quadratsumme, Summe der Quadrate der Totalen Abweichungen (SQT):
 - ▶ erfasst die „Gesamtvariation“ in der abhängigen Variablen
 - ▶ $SQT = \sum_{i=1}^n (Y_i - \bar{Y})^2$
- ▶ Residuenquadratsumme (SQR) :
 - ▶ beschreibt die Ungenauigkeit des Modells
 - ▶ $SQR = \sum_{i=1}^n (Y_i - \hat{Y})^2$

www.wikipedia.org

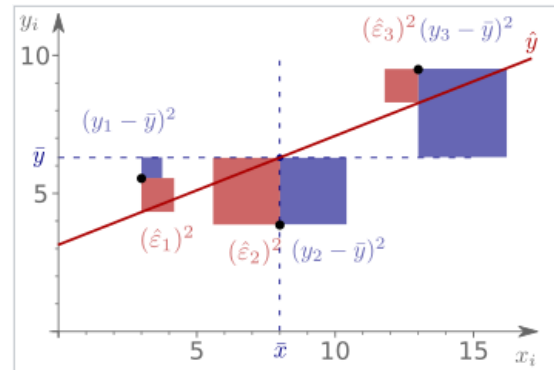


Bestimmtheitsmaß (R^2 , r^2)

- engl. Coefficient of Determination
- y : Beobachtungen, \hat{y}_i Vorhersagen

$$R^2 = 1 - \frac{\text{Residuenquadratsumme}}{\text{totalen Quadratsumme}}$$

- Wie viel Streuung in den Daten durch ein lineares Regressions-model „erklärt“ werden kann
- $R(0, 1)$
- Für einfache lineare Regression
 $r^2 = \text{Korrelationskoeffizient}^2$



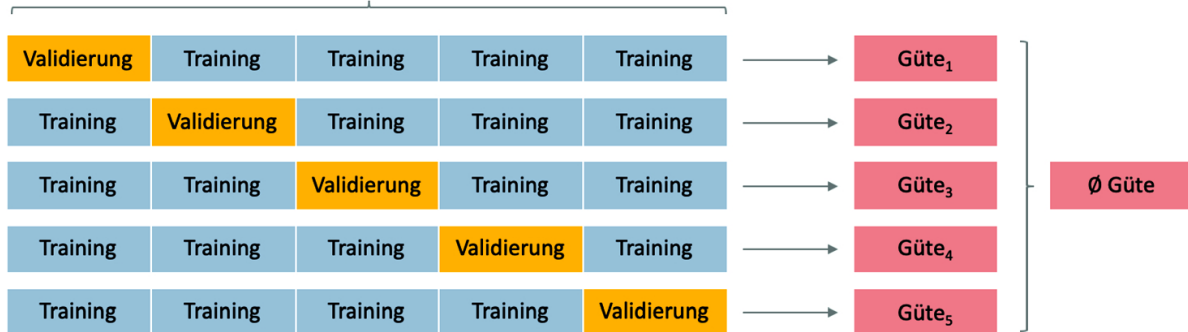
www.wikipedia.org

Kreuzvalidierung (cross validation)

- ▶ Unterteilung in „Trainingsdaten “ und „Testdaten“
- ▶ Regression mit den Trainingsdaten
- ▶ Vergleich der Regressionsergebnisse mit den Testdaten
- ▶ Bewertung der Güte der Regression
- ▶ iterative Analyse mit verschiedenen Trainings-/Testdatensätzen

Kreuzvalidierung mit $k=5$ Partitionen

Schacht & Lanquillon (2019)



Annahmen für lineare Regression

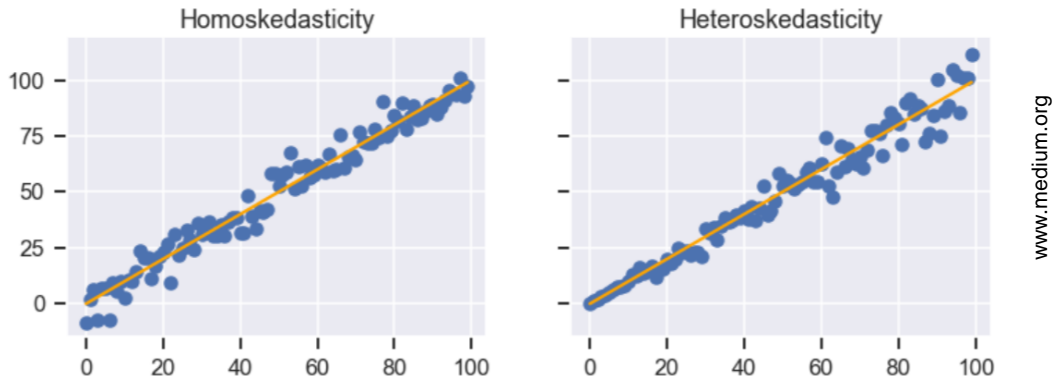
- ▶ Abhängige Variable ist eine Linearkombination der Regressionskoeffizienten
 - ▶ aber nicht zwingend der unabhängigen Variablen
 - ▶ Transformation der Daten

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- ▶ Normalverteilung der unabhängigen Variablen
 - ▶ Verallgemeinerte lineare Modelle (generalized linear models)
 - ▶ Verteilungen aus der Exponentialfamilie (Poisson, Gamma, usw.)
 - ▶ Diskrete Variablen → logistische Regression (nächste Stunde)

Annahmen für KQ-Schätzung

- ▶ Residuen sind normalverteilt $\sim (0, \sigma)$, homoskedastisch und weisen keine Autokorrelation auf
- ▶ Tests für Homoskedastizität: z.B. Breusch-Pagan, White test, ...
- ▶ Alternative: Verallgemeinerte KQ-Schätzung (weighted least squares)
- ▶ Berechnung gewichtete Residuen-Quadratsumme



www.medium.org

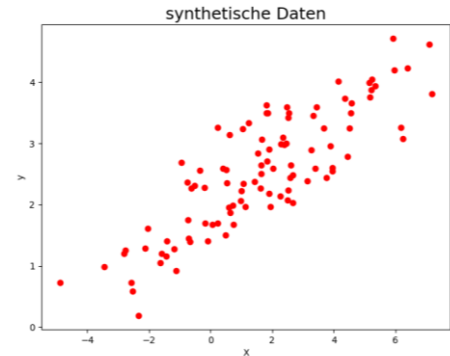
Multikollinearität

- ▶ Korrelation von zwei oder mehr unabhängigen Variablen
- ▶ KQ-Schätzung wird ineffizient und ungenau
 - ▶ Hohe Varianz im Regressionsmodell
 - ▶ Hohes Bestimmtheitsmaß R^2
- ▶ Identifikation über Korrelationsmatrix
- ▶ gilt für lineare und verallgemeinerte Regressionsmodelle



Übung 2.13: Lineare Regression

- ▶ Lineare Regression in Python
 - ▶ Einfache lineare Regression „from scratch“
 - ▶ Multiple lineare Regression mit scikit-learn
 - ▶ Fehlermaße
 - ▶ Validierung mit Hilfe von Trainings- und Test-Daten
- ▶ Aufgaben in Jupyter Notebook: geodatenanalyse_1-2-13

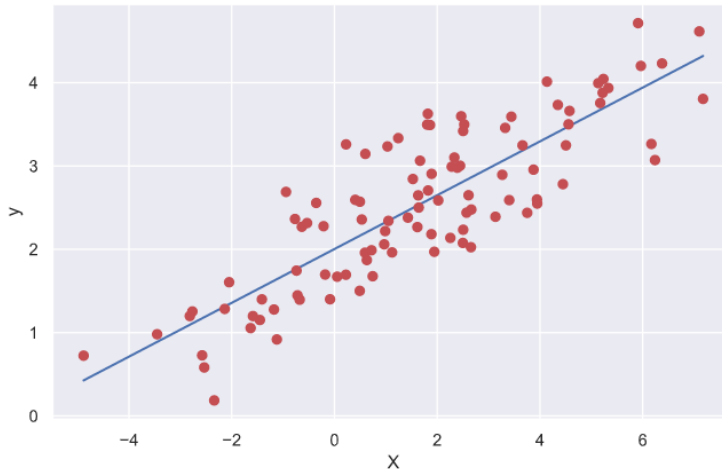


sharemomentssharelife/Flickr

Aufgabenbesprechung

- ▶ Einfache, lineare Regression „from scratch“
- ▶ Übereinstimmung der Regressionskoeffizienten

Beobachtet vs Vorhergesagt



```
print (alpha, beta)
```

```
2.0031670124623426 0.3229396867092763
```

```
np.random.seed(0)
```

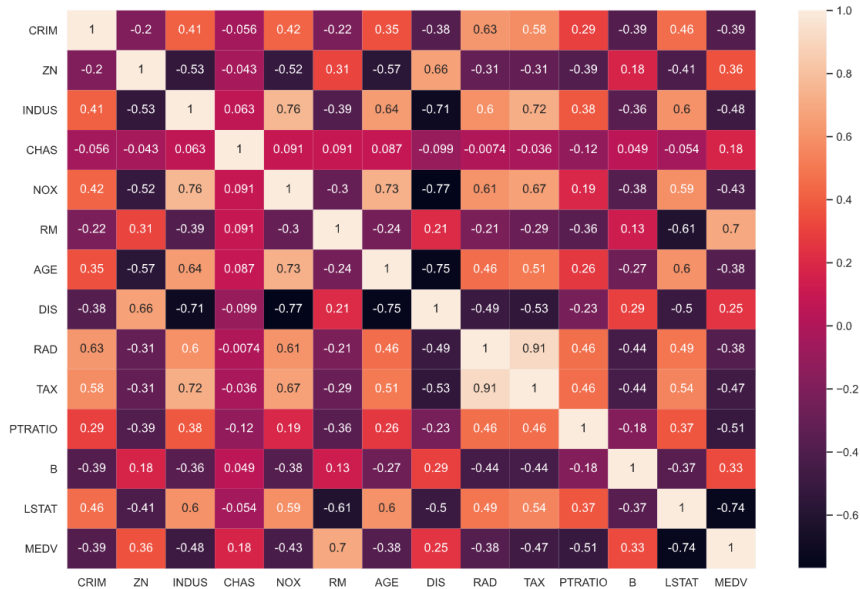
```
X = 2.5 * np.random.randn(100) + 1.5
```

```
res = 0.5 * np.random.randn(100)
```

```
y = 2 + 0.3 * X + res
```

Aufgabenbesprechung

- Multiple lineare Regression mit scikit-learn
- Parameterauswahl: „LSTAT“ und „RM“



Modell Evaluation Trainingsdaten:
5.6371293350711955 0.6300745149331701
Modell Evaluation Testdaten:
5.13740078470291 0.6628996975186954

Literatur

- ▶ Trauth (2015): MATLAB Recipes for Earth Sciences (4th Ed.), Springer
- ▶ Gelman et al. (2020) Regression and Other Stories, Cambridge University Press

Nützliche Weblinks:

- ▶ <https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0>

