

Geodatenanalyse I: Regressionsanalyse – Lineare Regression

Kathrin Menberg



Stundenplan

Vorläufiger Stundenplan		
Datum	Thema	Dozent
20.10.2021	Einführung in die Programmierung mit <i>Python</i>	Gabriel Rau
25.10.2021	Univariate Statistik und statistisches Testen	Kathrin Menberg
01.11.2021	<i>Feiertag</i>	
08.11.2021	Umgang und Berechnung von Datensätzen	Gabriel Rau
15.11.2021	Bivariate und schließende Statistik	Kathrin Menberg
22.11.2021	Datenvisualisierung mit <i>matplotlib</i>	Gabriel Rau
29.11.2021	Multivariate Statistik	Kathrin Menberg
06.12.2021	Datenformate, Datenspeicherung und Datenbanken	Gabriel Rau
13.12.2021	Monte-Carlo Methoden	Kathrin Menberg
20.12.2021	Analyse und Visualisierung von Geodaten	Gabriel Rau
27.12.2021	<i>Weihnachtsferien</i>	
03.01.2022	<i>Weihnachtsferien</i>	
10.01.2022	Sensitivitätsanalyse	Kathrin Menberg
17.01.2022	Datenethik, Lizenzierung und Entwicklungstools	Gabriel Rau
24.01.2022	Räumliche Interpolation	Kathrin Menberg
31.01.2022	Fragen zur Programmierung	Gabriel Rau
07.02.2022	Regressionsanalyse	Kathrin Menberg

Vorlesungsplan

Uhrzeit	Inhalt
10:00 – 10:30	Regressionsanalyse
10:30 – 11:15	Übung
11:15 – 11:30	<u>Pause</u>
11:30 – 12:15	Fortsetzung Übung
12:15 – 12:30	Diskussion und Reflexion

6339042: Geodatenanalyse I – Programmierung und Geostatistik

Prüfungsleistung

► Prüfungsaufgabe

- Bearbeitung einer vorgegebenen Aufgabenstellung in *Python*
- Erstellen eines individuellen Workflows mit Code und Erklärung zur Analyse eines Geodatensatzes
- Dokumentation in Form eines Jupyter Notebooks mit Visualisierung und Diskussion der Ergebnisse
- Abgabe bis 31.05.2022
- Die Prüfungsaufgabe wird benotet und entspricht der Modulnote
- Für die Prüfungsaufgabe sind ca. 60 Stunden Arbeit veranschlagt

Lernziele

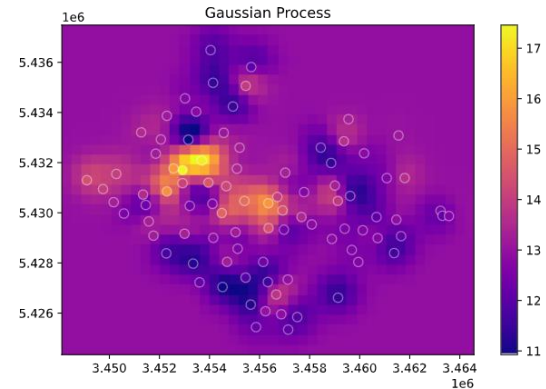
Am Ende der Stunde werden die Teilnehmer:

- ▶ ... 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.

Wozu Regressionsanalyse?

► Vorhersagen (prediction)

- Modellierung von existierenden Beobachtungen
- Neue Datenwerte vorhersagen
- Siehe Interpolation mit Kriging



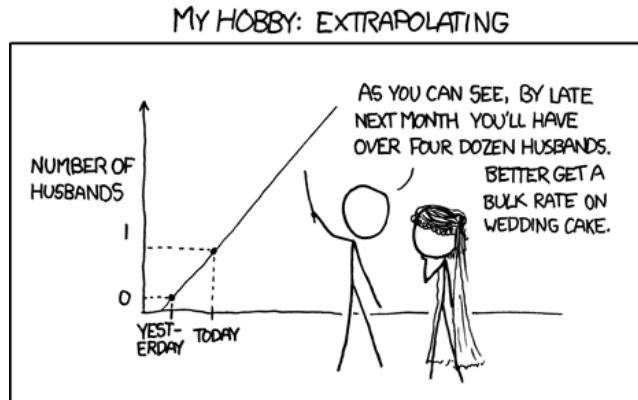
► Variablenassoziation

- Zusammenhänge von Variablen identifizieren
- Gliederungen und Strukturen in Datensätzen

Wozu Regressionsanalyse?

► Extrapolation

- Ausgleichen des Unterschieds zwischen Stichprobe und Grundgesamtheit

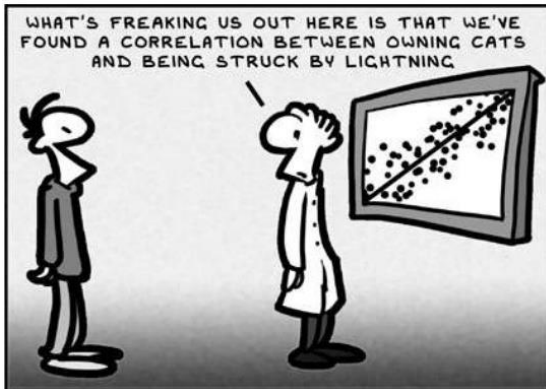


www.pinterest.at

► Kausale Schlussfolgerungen

(causal inference)

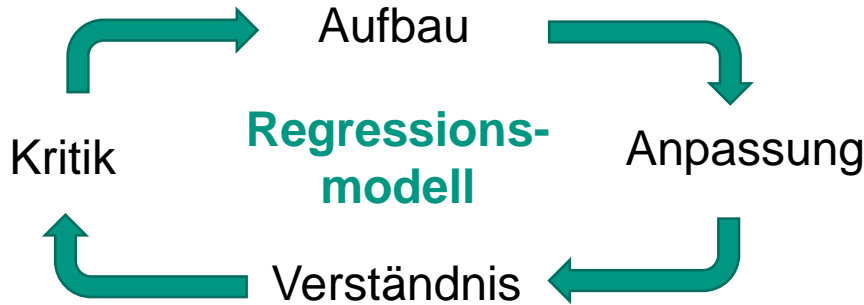
- Effekte von Verfahren (Variablenänderungen) ableiten
- Experimentelles Design!



4-Stufen Zyklus der statistischen Analyse

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

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

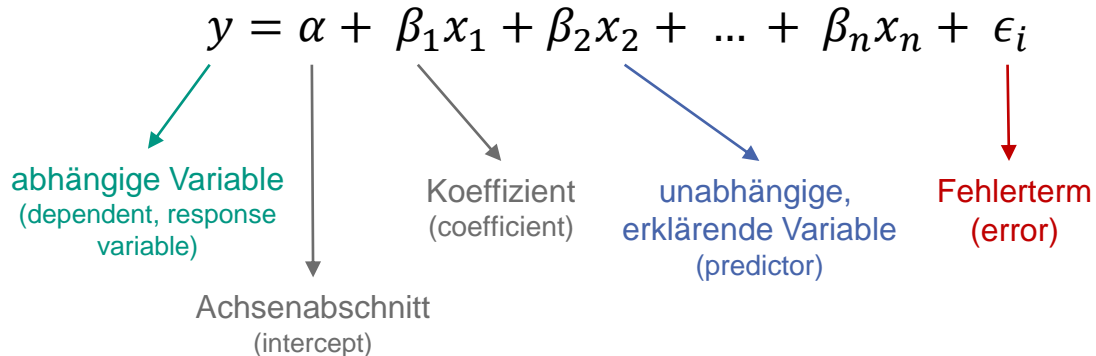


- ▶ Graphische Darstellung
- ▶ Beziehungen zwischen Variablen und Messungen untersuchen

- ▶ Datenmanipulation
- ▶ Koeffizienten schätzen
- ▶ Unsicherheiten

Grundlagen lineare Regression

- ▶ Abhängige Variable als eine Linearkombination der Regressionskoeffizienten

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


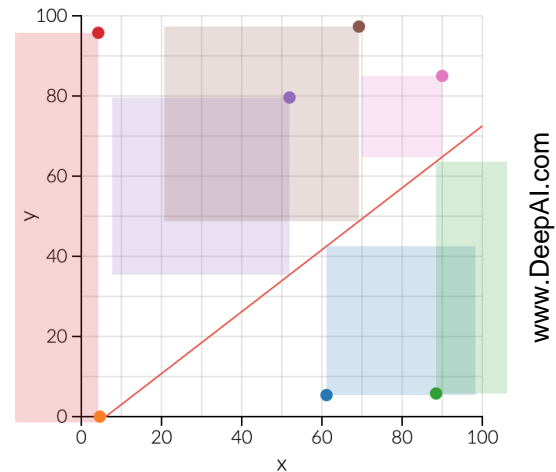
The diagram illustrates the components of the linear regression equation $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$. Arrows point from each term to its description:

- 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 $\hat{y} = (\alpha, \beta)$



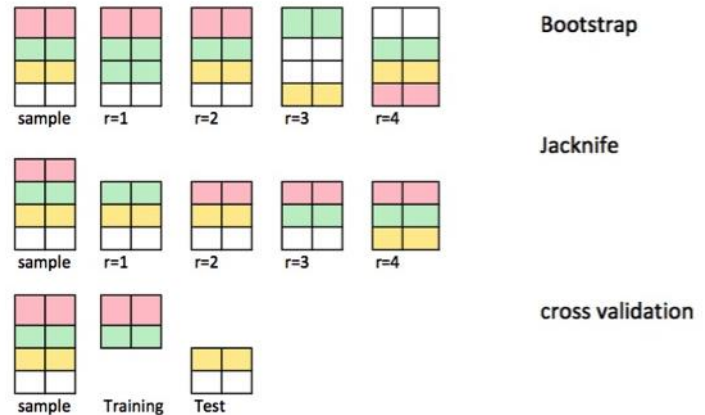
Überprüfung der Anpassungsgüte

► Fehlermaße:

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

► Methoden zur Validierung:

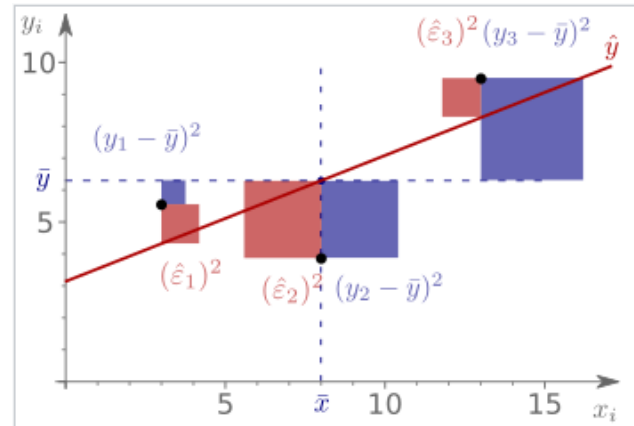
- Bootstrap und Jackknife
- Kreuzvalidierung
- u.v.m.



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

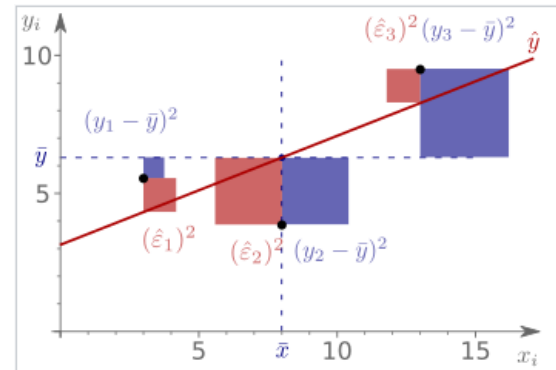


Bestimmtheitsmaß (R^2 , r^2)

- 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
 $r^2 = \text{Korrelationskoeffizient}^2$

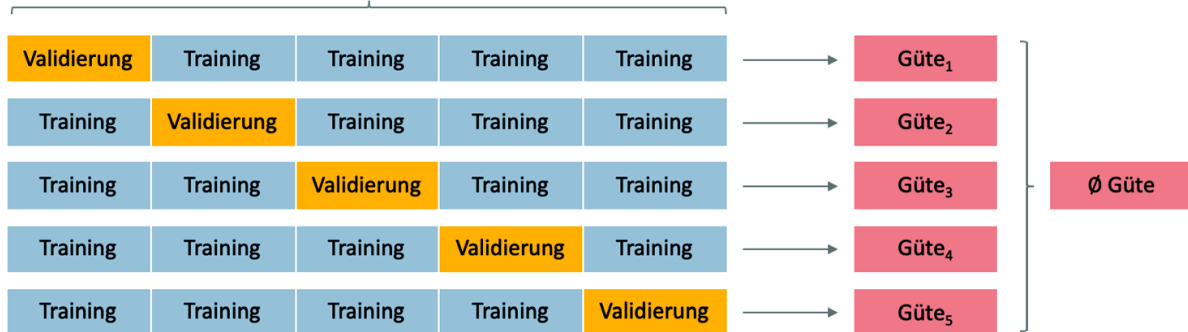


www.wikipedia.org

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



Annahmen für lineare Regression

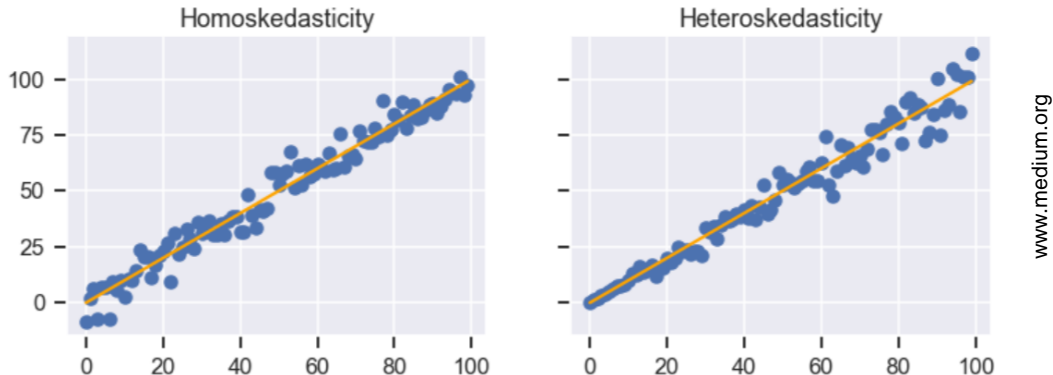
- ▶ Abhängige Variable ist eine Linearkombination der Regressionskoeffizienten
 - ▶ aber nicht zwingend der unabhängigen Variablen
 - ▶ Transformation der Daten

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

- ▶ 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



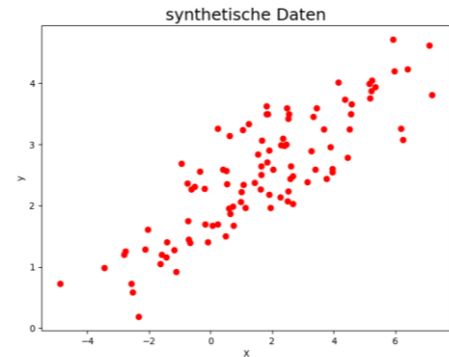
Multikollinearität

- ▶ 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 11: Lineare Regression

- ▶ Lineare Regression in Python
 - ▶ Multiple lineare Regression mit scikit-learn
 - ▶ Fehlermaße
 - ▶ Validierung mit Hilfe von Trainings- und Test-Daten
- ▶ Aufgaben in Jupyter Notebook:
11_Lineare Regression_uebung



sharemomentssharelife/Flickr

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>

