Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Regressionsanalyse

Violeta Vogel, TI BFH

Regression Definition

- Ein statistisches Analyseverfahren
- Ziele:
  - Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen im gegebenen Datensatz zu modellieren.
  - Zusammenhänge quantitativ zu beschreiben
  - Werte der abhängigen Variablen zu prognostizieren.

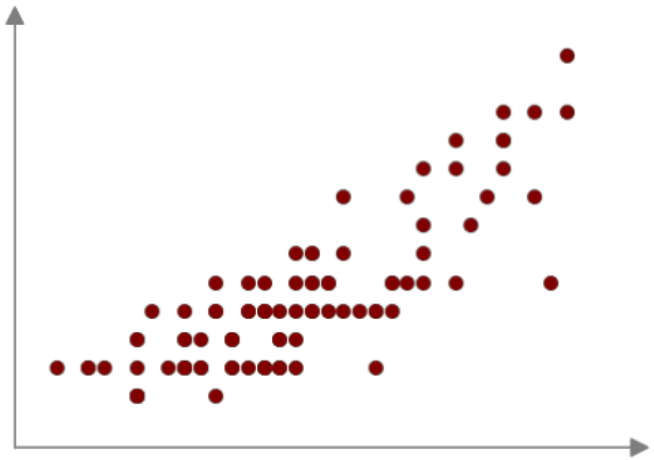
Typen der Regression

- Einfache lineare Regression: Vorhersage einer Variablen auf Basis einer anderen Variablen
- Multiple lineare Regression: mehrere unabhängige Faktoren eine abhängige Variable beeinflussen können und somit auch für ihre Vorhersage in Betracht kommen
- Logistische Regression: Wird verwendet, wenn die abhängige Variable kategorial ist, z.B. binär (ja/nein) oder multinominal (mehrere Kategorien).
- Ridge Regression, Lasso Regression: Sind spezielle Formen der linearen Regression, die bei multicollinearität (hohe Korrelation zwischen den Prädiktoren) eingesetzt werden.

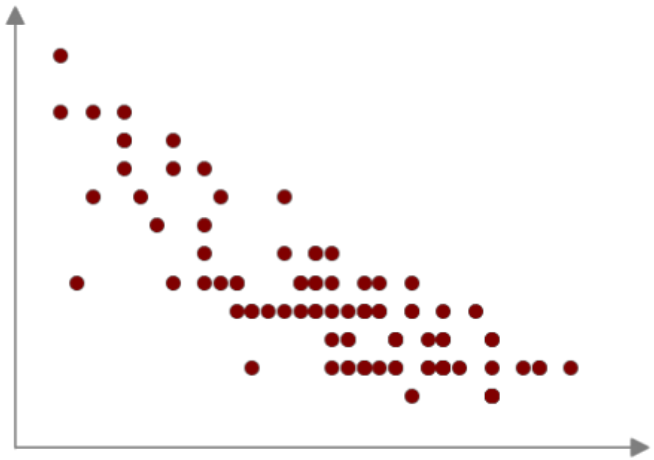
Korrelation

- Korrelation beschreibt wie Variablen zusammenhängen:
  - Positiver Zusammenhang: je höher der Wert einer Variable, desto höher der Wert der anderen Variable. Je niedriger der Werte einer Variable, desto niedriger der Wert der anderen Variable.
  - Negativer Zusammenhang: je höher der Wert einer Variable, desto niedriger der Wert der anderen Variable. Je niedriger der Wert der einen Variable, desto höher der Wert der anderen Variable
  - Kein Zusammenhang: die Höhe der Werte auf beiden Variablen variieren nicht miteinander. Eine Veränderung der einen Variable hat keinen Einfluss auf die Veränderung der anderen Variable.

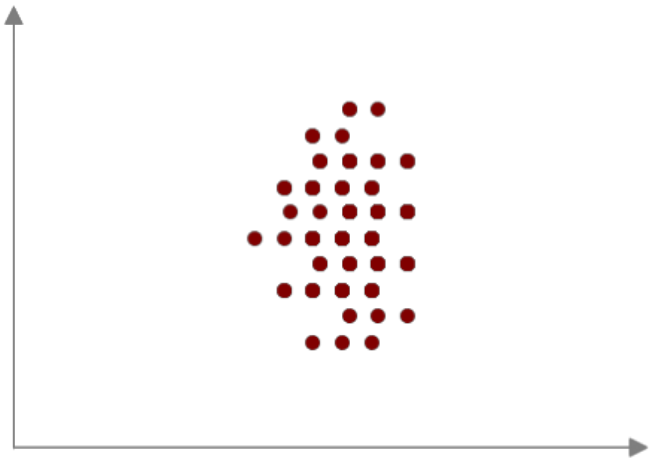
Korrelation linearer Zusammenhang



Positiver Zusammenhang

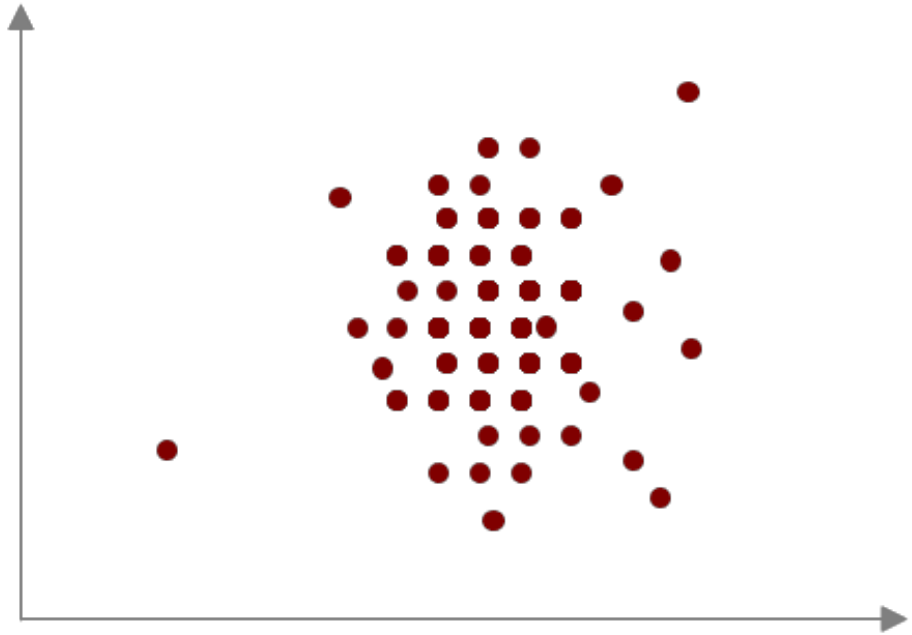


Negativer Zusammenhang

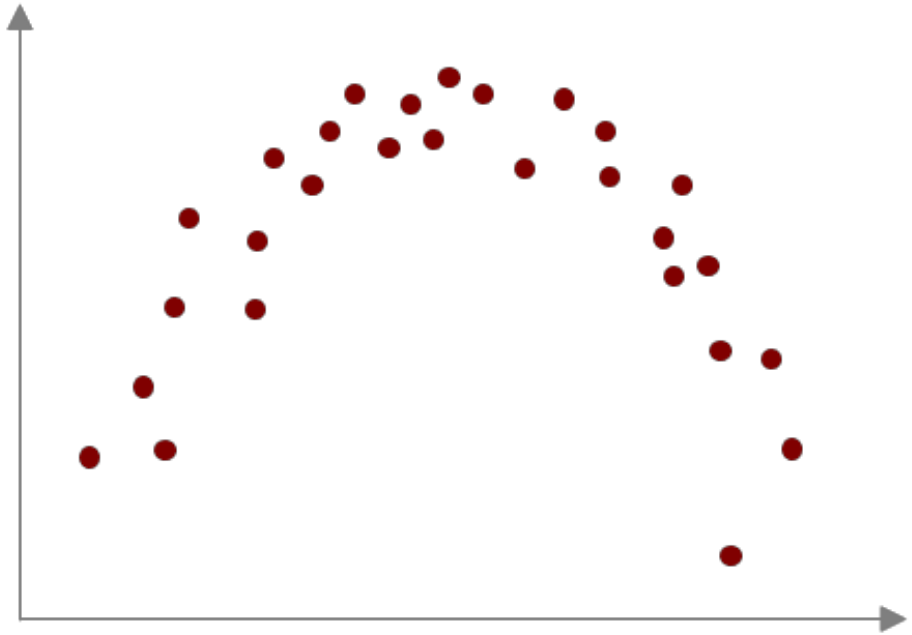


Kein Zusammenhang

Korrelation: Achtung nicht linearer Zusammenhang



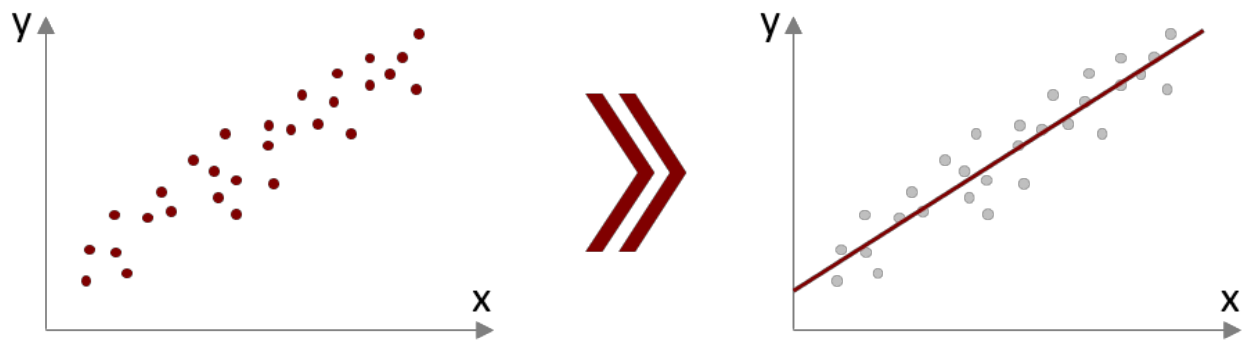
$r \approx 0$



$r \approx 0$

Korrelation und Regression

- Die Korrelation beschäftigt sich mit der Frage nach dem Zusammenhang zwischen zwei Variablen.
- Die Regression nutzt diesen Zusammenhang, um Werte der einen Variable auf Basis der Werte der anderen Variable vorherzusagen



r = .8

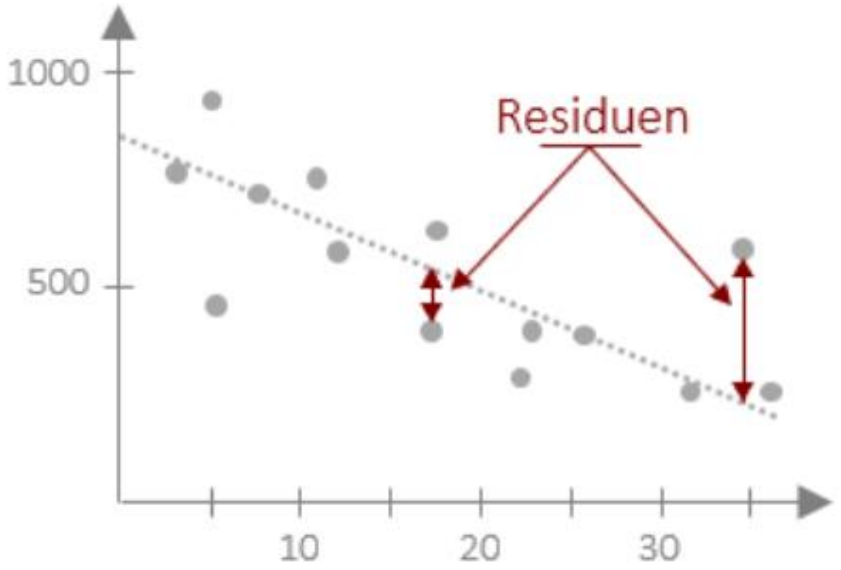
Korrelation

y = 1 + 0,5*x*

Regression

Residuum

- Abstand aller Punkte zur Linie soll minimal sein. Der Abstand der Werte zur Vorhersagelinie wird auch Residuum (Rest) genannt
- Wird mit der Methode der kleinsten Quadrate bestimmt



Regression: Wahl der Methode

- Die Wahl der passenden Regressionsart hängt von der Art der Daten und der Fragestellung der Analyse ab.

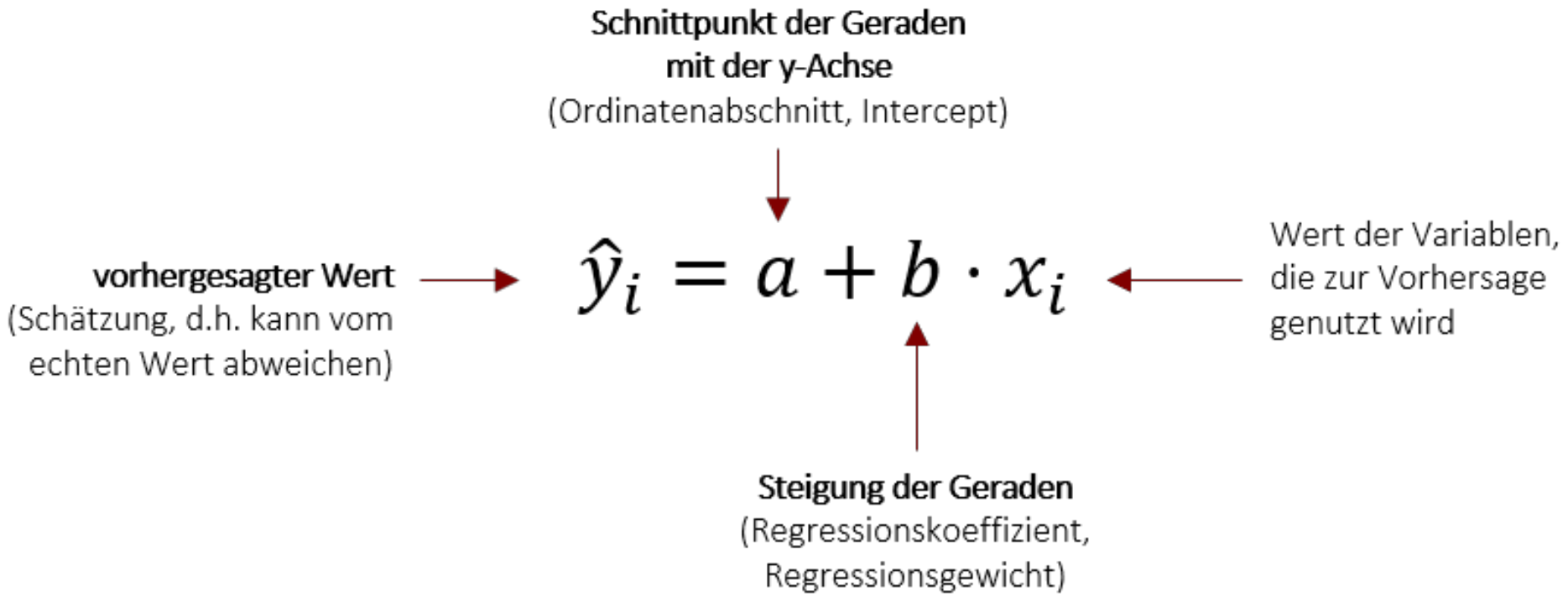
Regression je nach Datentyp

Art der Regression	Skalentyp der abhängigen Variable (AV) oder Kriterium	Skalentyp der unabhängigen Variable (UV) oder Prädiktor
Einfache lineare Regression	Metrisch	Metrisch
Multiple lineare Regression	Metrisch	Metrisch Ordinal Dichotom
Logistische Regression	Dichotom Intervallskaliert Diskret	Beliebig
Multinominale logistische Regression	Kategorial (binär oder multinominal (mehrere Kategorien))	Beliebig

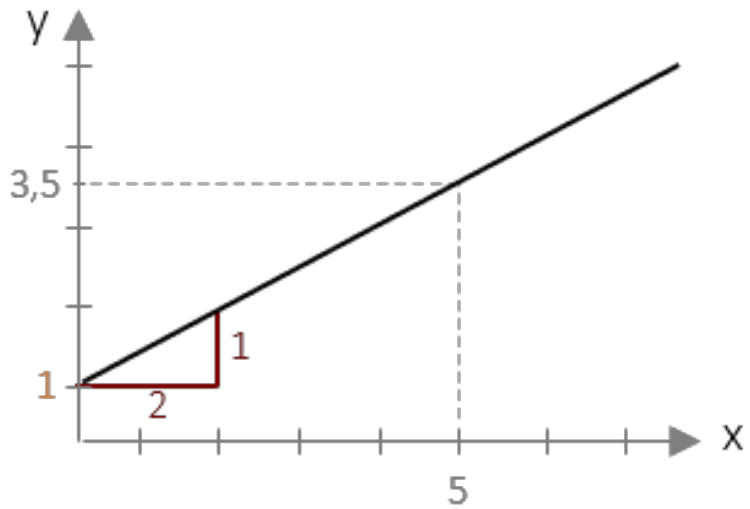
Lineare Regression

- Regression – Statistik Grundlagen
- Mit einer einfachen linearen Regression wird die lineare Beziehung zwischen zwei stetigen Variablen untersucht

Regressionsgleichung



Regressionsgleichung



Regressionsgleichung

$$\hat{y} = 0{,}5x + 1$$

Beispiel für x=5

$$3{,}5 = 0{,}5 \cdot 5 + 1$$

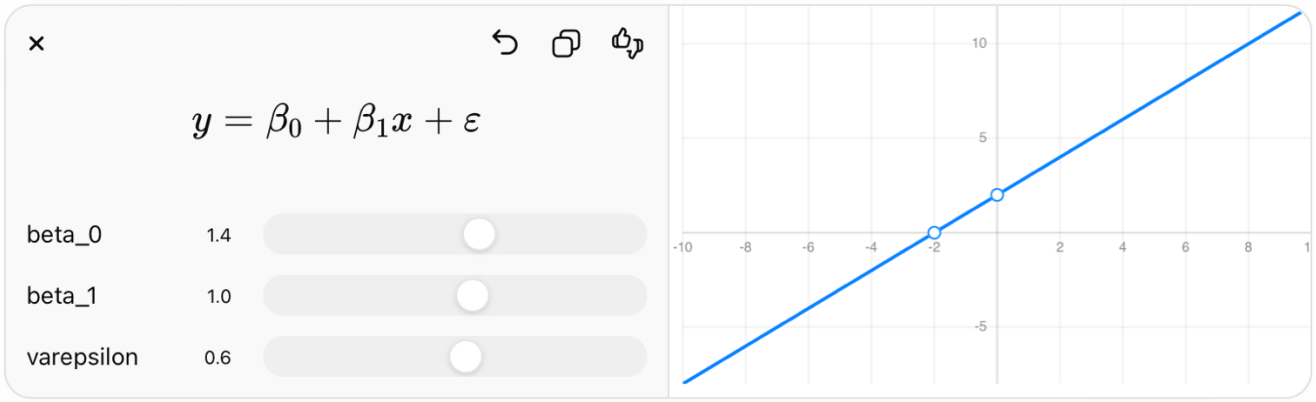
Voraussetzung für die lineare Regression

- Zwischen den Variablen besteht ein linearer Zusammenhang.
- Das Skalenniveau der AV und UV sollte metrisch sein, sprich einen konkreten Zahlenwert besitzen. Ein Beispiel dafür ist die Körpergröße.
- Die Residuen (Abweichungen) sollten zum einen keine Korrelation untereinander aufweisen und zum anderen konstant über den gesamten Wertebereich der AV streuen. Dies wird Homoskedastizität genannt.
- Es sollten möglichst wenig Ausreiser in den Daten sein, da diese einen großen Einfluss auf die Vorhersagegüte haben können.

Lineare Regression: Beispiel

- Wie beeinflusst die Lernzeit die Prüfungsnote?
- y: Prüfungsnote
- x: Lernzeit
- β0: Achsenabschnitt
- β1: Einfluss der Lernze
- ε: Fehlerterm

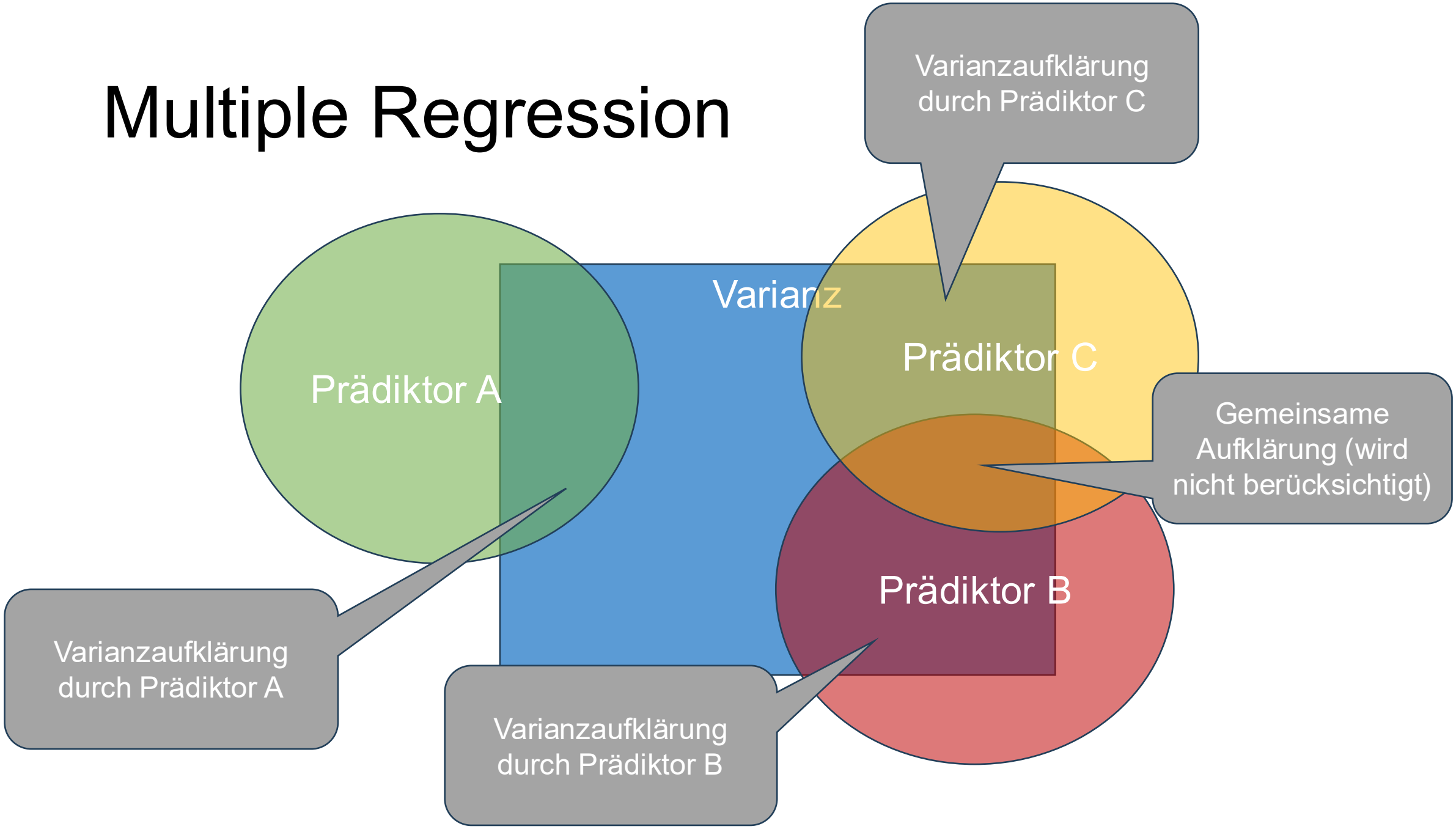
Modell:



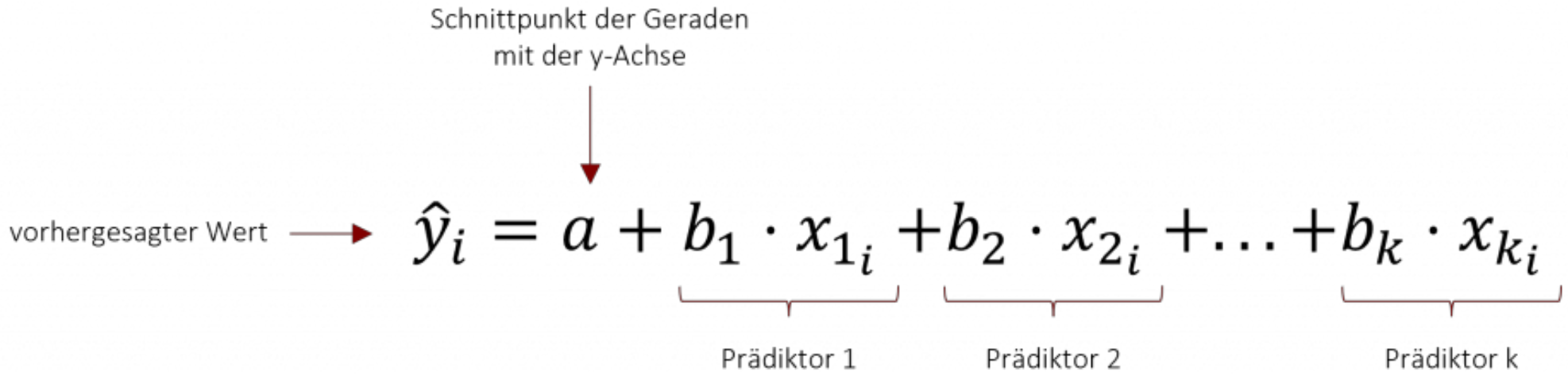
Multiple Regression

- Multiple Regression – Statistik Grundlagen
  - Bei der multiplen linearen Regression werden die linearen Beziehungen zwischen einer stetigen Antwortvariablen und mindestens zwei Prädiktoren untersucht.
  - Bei der multiplen Regression gibt es mehrere unabhängige Variablen, die gemeinsam y erklären.

Multiple Regression



Regressionsgleichung multiple Regression



Multiple lineare Regression: Regressionsgewicht

	Wertebereich	Eigenschaften
Unstandardisiertes Regressionsgewicht b	Abhängig vor der gewählten Skala	• Inhaltlich interpretierbar • Zwischen Prädiktoren vergleichbar, wenn du diese auf selben Skala gemessen werden
Standardisiertes Regressionsgewicht β (an der Standardabweichung)	[-1, +1]	• Inhaltlich nicht interpretierbar • Zwischen Prädiktoren vergleichbar • Interpretierbar Korrelationen (r)

Multiple lineare Regression: Regressionsgewicht

	b	β	T	Sig.
(Konstante)	0,89	0,72		
Geschmack	1,98	0,414	3,576	0.001 **
Aussehen	0,32	0,049	0,387	0,682
Preis	0,76	0,541	4,841	0.000 ***

- Erste Zeile: Die Konstante spiegelt dabei den Ordinatenabschnitt wider
- Zweite Spalte: Die standardisierten Regressionsgewichte β
- Dritte Spalte: Hat der jeweilige Prädiktor einen signifikanten Einfluss auf die abhängige Variable hat
- Vierte Spalte: Sollte man von einem Zufall ausgehen

Voraussetzungen für die multiple lineare Regression

- Das Skalenniveau der AV und UV muss metrisch sein
- Linearer Zusammenhang zwischen den Prädiktoren und der abhängigen Variable
  - (Überprüfung graphisch)
- Die Residuen sollen normalverteilt sein
  - (Bei einer ausreichend großen Stichprobengröße kann die Überprüfung dieser Voraussetzung vernachlässigt werden)
- Homoskedastizität: Streuungen der zu einem x-Wert gehörenden y-Werte müssen über den ganzen Wertebereich von x homogen sein
  - (Levene Test oder graphisch)
- Ausreiser oder einflussreiche Punkte
  - (Leverage-Analyse)
- Residuen dürfen nicht korrelieren
  - Dies ist nicht weiter problematisch, solange die Korrelationen zwischen den Prädiktoren einen Schwellenwert von r = 0.8 nicht überschreiten.

Multiple lineare Regression: Beispiel

- Wie werden Prüfungsnoten durch Lernzeit, Unterrichtsbesuche und Schlafdauer beeinflusst?
- Modell:
  - $y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta kxk+\varepsilon$
  - x_1: Lernzeit
  - x_2: Unterrichtsbesuche
  - x_3: Schlafdauer
  - usw.
  - Jeder Koeffizient β_i zeigt den Einfluss der jeweiligen Variable auf y, während die anderen Variablen konstant gehalten werden.

Multiple lineare Regression: Merksatz

- Jede multiple Regression ist eine lineare Regression (weil sie linear in den Koeffizienten ist), aber nicht jede lineare Regression ist eine multiple Regression.
- Die multiple Regression erweitert die einfache lineare Regression um mehrere Einflussgrössen.

Multiple vs. lineare Regression

Merkmal	Lineare Regression	Multiple Regression
Anzahl unabhängiger Variablen	1	2 oder mehr
Ziel	Zusammenhang zwischen x und y untersuchen	Gemeinsamen Einfluss mehrerer Variablen auf $y$untersuchen
Beispiel	Umsatz durch Werbebudget erklären	Umsatz durch Werbebudget, Preis und Saison erklären
Grafische Darstellung	Gerade in 2D	Ebene bzw. Hyperfläche in mehreren Dimensionen

Use Cases für Regression

- Nennen Sie ein paar konkrete Fragestellungen aus folgenden Gebieten:
  - Immobilienpreise
  - Gehaltsanalyse
  - Autopreise
  - Medizin
  - Sportanalyse
  - Finanzdaten

Use Cases für Regression

- Was können Unabhängige Variablen sein, was abhängige, was das Ziel?
- Nennen Sie ein paar konkrete Fragestellungen aus folgenden Gebieten:
  - Immobilienpreise
  - Gehaltsanalyse
  - Autopreise
  - Medizin
  - Sportanalyse
  - Finanzdaten

Regression: Herausforderungen

- Nichtlineare Zusammenhänge
  - Die einfache lineare Regression geht davon aus, dass zwischen Eingangsvariablen und Zielvariable ein linearer Zusammenhang besteht. In der Realität sind Zusammenhänge oft komplexer.
  - Beispiel: Der Einfluss von Alter auf Einkommen ist selten streng linear.
- Multikollinearität
  - Wenn mehrere Einflussgrössen stark miteinander korrelieren, wird es schwierig, ihre einzelnen Effekte zuverlässig zu bestimmen.
  - Beispiel: Wohnfläche und Anzahl der Zimmer bei Immobilienpreisen.
- Ausreisser (Outlier)
  - Einige wenige extreme Beobachtungen können die Regressionsgerade stark verzerren und zu schlechten Schätzungen führen.

Regression: Herausforderungen

- Overfitting und Underfitting
  - Overfitting: Das Modell passt sich zu stark an die Trainingsdaten an und generalisiert schlecht.
  - Underfitting: Das Modell ist zu einfach und erfasst wichtige Muster nicht.
- Heteroskedastizität
  - Die Streuung der Fehler ist nicht über alle Werte hinweg konstant. Dadurch können Schätzungen und Tests unzuverlässig werden.
- Fehlende Daten
  - Unvollständige Datensätze können die Analyse erschweren und zu verzerrten Ergebnissen führen.

Regression: Herausforderungen

- Verletzung der Modellannahmen
  - Bei der linearen Regression werden oft folgende Annahmen getroffen:
    - Linearität
    - Unabhängigkeit der Fehler
    - Konstante Fehlervarianz
    - Normalverteilung der Residuen
  - Werden diese verletzt, sinkt die Aussagekraft des Modells.
- Auswahl relevanter Variablen
  - Zu viele irrelevante Variablen erhöhen die Komplexität, während wichtige fehlende Variablen zu verzerrten Ergebnissen führen können.

Regression: Herausforderungen

- Kausalität vs. Korrelation
  - Regression kann Zusammenhänge aufzeigen, beweist aber normalerweise keine Ursache-Wirkungs-Beziehung.
  - Beispiel: Eisverkauf und Badeunfälle steigen beide im Sommer, aber Eis verursacht keine Badeunfälle.
- Interpretierbarkeit
  - Komplexere Regressionsmodelle können zwar genauer sein, sind aber oft schwieriger zu interpretieren als einfache lineare Modelle.

Lasso Regression

- Lasso-Regression (L1-Regularisierung) ist eine Technik in der Regressionsanalyse, die sowohl eine Variablenselektion als auch eine Regularisierung durchführt
  - Ziel: Überanpassung zu verhindern und die Genauigkeit von Modellen zu verbessern
  - Vorgehen: einige Koeffizienten der unabhängigen Variablen werden auf Null gesetzt.
- Ergebnis: Spärlichen Modelle

Lasso Regression Formel

$$L = \text{Summe}[(y - (a + bx))^2] + \lambda * \text{Summe}[|b|]$$

- ‘y’ für die tatsächlichen Werte,
- ‘a’ für den Y-Achsenabschnitt,
- ‘b’ für die Steigung oder den Koeffizienten
- ‘x’ für die Vorhersagewerte.
- ‘λ’ ist der Regularisierungsparameter, der steuert, wie stark die Koeffizienten geschrumpft werden. Wenn λ = 0 ist, haben wir eine normale lineare Regression, und wenn λ sehr groß ist, werden alle Koeffizienten auf Null gesetzt

Regularisierungsparameter Lambda

- Regularisierungsparameter steuert den Grad der angewendeten Regularisierung:
  - Grössere Lambda-Werte erhöhen diese Strafe, wodurch mehr Koeffizienten gegen Null gehen. Dadurch werden einige der Merkmale des Modells weniger wichtig (oder sogar ganz eliminiert), was zu einer automatischen Auswahl der Merkmale führt
  - Kleinere Lambda-Werte verringern die Auswirkung der Strafe, sodass mehr Merkmale im Modell erhalten bleiben
  - Wenn λ gleich null ist, bleibt eine OLS-Funktion übrig, d. h. ein Standard-Lineares-Regressionsmodell ohne Regularisierung

Lasso Regression

- Regularisierung: Lasso-Regression fügt der Zielfunktion einen Strafterm hinzu, der die Summe der absoluten Werte der Koeffizienten (L1-Norm) multipliziert mit einem Regularisierungsparameter (λ) beinhaltet.
- Variablenselektion: Einige Koeffizienten werden auf Null gesetzt, was bedeutet, dass die entsprechenden Variablen aus dem Modell entfernt werden
- Überanpassung verhindern: Durch die Regularisierung wird die Komplexität des Modells reduziert, was dazu beiträgt, Überanpassung zu vermeiden
- Sparsität: Lasso-Regression führt zu spärlichen Modellen

Einsatz Lasso Regression

- Genomische Auswahl:
  - In der Genetik: die Anzahl der Variablen (Gene) zu reduzieren, die in einem Modell berücksichtigt werden müssen. Ergebnis: Identifikation der Gene, die am stärksten mit bestimmten Krankheiten oder Merkmalen korrelieren.
- Vorhersage in Wirtschaft und Finanzen:
  - Vorhersage der zukünftigen wirtschaftlichen und finanziellen Trends
  - Vorhersage von Aktienkursen auf der Grundlage verschiedener wirtschaftlicher Indikatoren verwendet werden.
- Kreditrisikobewertung:
  - Banken und andere Finanzinstitutionen: Bewertung des Risikos, dass ein Kreditnehmer seinen Kredit nicht zurückzahlt.

Nachteile Lasso Regression

- Auswahl des Alpha-Parameters:
  - Ein zu kleiner Alpha-Wert kann dazu führen, dass das Modell Overfitting aufweist
  - Ein zu großer Alpha-Wert kann dazu führen, dass das Modell zu einfach wird und an Vorhersagegenauigkeit verliert.
  - Es kann einige Experimente erfordern, um den richtigen Alpha-Wert zu finden
- Leistung bei korrelierten Variablen:
  - kann bei stark korrelierten Variablen schlechter abschneiden, indem nur eine der korrelierten Variablen auswählt und die anderen ignoriert
  - Ergebnis: Verlust der Informationen
- Nicht geeignet für alle Datensätze

Nachteile Lasso Regression

Vorteile	Nachteile
Automatische Variablenselektion	Problematisch bei stark korrelierten Variablen
Verhindert Overfitting	Wahl von λ kann schwierig sein
Einfachere Interpretation	Kann wichtige Variablen entfernen
Geeignet für viele Merkmale	Erfasst Nichtlinearitäten nur begrenzt

Ridge Regression (L2-Regularisierung)

- Eine Abwandlung der linearen Regression,
  - die um einen zusätzlichen Regularisierungsterm erweitert wurde,
  - sodass eine Überanpassung vermieden werden soll
- berücksichtigt zusätzlich noch die Größe der Regressionskoeffizienten
- versucht, einzelne, sehr große Modellkoeffizienten zu verhindern
- Folge: Die Wahrscheinlichkeit ist geringer, dass das Modell übermäßig komplexe Beziehungen aus den Trainingsdaten erlernt und dadurch overfitted

Ridge Regression Formel

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Dabei gilt:

- y_i: tatsächlicher Wert
- $\hat{y}_i$: vorhergesagter Wert
- β_j: Regressionskoeffizienten
- λ: Regularisierungsparameter

Je grösser λ, desto stärker werden die Koeffizienten verkleinert.

Ridge Regression Formel

- Kostenfunktion ist die ‘beste’ Linie ist, die die Summe der quadratischen Differenzen (oder “Fehler”) zwischen den tatsächlichen und den vorhergesagten Werten der Zielvariable minimiert
- Ridge Regression fügt dieser Kostenfunktion die “Regularisierung”
- Der Regularisierungsterm ist der Quadratwert der Koeffizienten (Gewichtungen, die wir unseren Eingangsvariablen geben), multipliziert mit einem Faktor “Alpha”
- Dieser Alpha-Wert ist ein Parameter:
  - ein hoher Wert bedeutet mehr Regularisierung
  - ein niedriger Wert bedeutet weniger Regularisierung
- Regularisierungsterm macht es ‘teurer’ für das Modell, grosse Koeffizienten zu haben:
  - Modell ist weniger komplex und weniger anfällig für Overfitting

Ridge Regression Herausforderungen

- Wahl des Regularisierungsparameters
  - Die Bestimmung des optimalen λ ist entscheidend.
  - Zu klein: kaum Unterschied zur linearen Regression.
  - Zu gross: wichtige Informationen gehen verloren.
- Keine Variablenselektion
  - Anders als Lasso entfernt Ridge keine Variablen vollständig. Dadurch kann das Modell schwieriger zu interpretieren sein.
- Standardisierung erforderlich
  - Da Variablen unterschiedliche Grössenordnungen haben können, sollten sie vor der Analyse standardisiert werden.

Ridge Regression Herausforderungen

- Nichtlineare Zusammenhänge
  - Die Ridge-Regression basiert auf einem linearen Modell und kann komplexe nichtlineare Beziehungen nur begrenzt abbilden.
- Interpretierbarkeit
  - Da alle Variablen im Modell verbleiben, kann die Interpretation bei vielen Merkmalen anspruchsvoll sein.

Ridge Regression Einsatz

- Immobilienbewertung
  - Vorhersage von Hauspreisen anhand vieler miteinander korrelierter Merkmale wie: Wohnfläche, Anzahl Zimmer, Grundstücksgrösse, Baujahr
- Finanzwesen: Prognose von Aktienrenditen oder Kreditrisiken bei vielen wirtschaftlichen Einflussgrössen.
- Marketing: Analyse des Einflusses verschiedener Marketingkanäle auf den Umsatz.
- Medizin: Vorhersage von Krankheitsrisiken anhand zahlreicher klinischer Merkmale.
- Machine Learning: Verbesserung der Generalisierungsfähigkeit von Modellen mit vielen Variablen.

Ridge Regression Vor- und Nachteile

Vorteile	Nachteile
Reduziert Overfitting	Keine automatische Variablenselektion
Gut bei Multikollinearität	Wahl von λ kann schwierig sein
Nutzt alle verfügbaren Informationen	Weniger gut interpretierbar bei vielen Variablen
Stabile Schätzungen	Erfasst Nichtlinearitäten nur begrenzt

Lasso versus Ridge

Ridge	Lasso
Verkleinert Koeffizienten	Verkleinert Koeffizienten
Setzt Koeffizienten selten auf null	Kann Koeffizienten auf null setzen
Behält alle Variablen	Führt Variablenselektion durch
Gut bei stark korrelierten Variablen	Wählt oft nur einige korrelierte Variablen aus

Evaluation Regressionen

1. Mean Absolute Error (MAE)

Der **Mean Absolute Error (MAE)** misst die durchschnittliche absolute Abweichung zwischen vorhergesagten und tatsächlichen Werten.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Vorteile

- Einfach zu interpretieren
- Weniger empfindlich gegenüber Ausreissern

Nachteil

- Grössere Fehler werden nicht besonders stark bestraft

Beispiel: Ein MAE von 5 bedeutet, dass die Vorhersagen durchschnittlich um 5 Einheiten vom tatsächlichen Wert abweichen.

Evaluation Regressionen

2. Mean Squared Error (MSE)

Der **Mean Squared Error (MSE)** berechnet die durchschnittliche quadratische Abweichung.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Vorteile

- Bestraft grosse Fehler stärker
- Weit verbreitet in Machine Learning

Nachteil

- Empfindlich gegenüber Ausreissern

Evaluation Regressionen

3. Root Mean Squared Error (RMSE)

Der **Root Mean Squared Error (RMSE)** ist die Quadratwurzel des MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Vorteile

- Gleiche Einheit wie die Zielvariable
- Grosse Fehler werden stärker gewichtet

Nachteil

- Sensibel gegenüber Ausreissern

Evaluation: Cross validation

- Cross Validation ist ein Verfahren zur Bewertung von Machine-Learning- und Regressionsmodellen.
- Ziel ist es zu überprüfen, wie gut ein Modell auf unbekannten Daten funktioniert und ob es gut generalisiert.
- Cross Validation reduziert dieses Risiko, indem mehrere Trainings- und Testaufteilungen verwendet werden
- Vorteile:
  - Zuverlässigere Bewertung der Modellleistung
  - Erkennung von Overfitting
  - Bessere Modellauswahl
  - Effiziente Nutzung der verfügbaren Daten

Evaluation: Cross validation

- Vorgehen
  - Datensatz in k gleich grosse Teile (Folds) aufteilen.
  - Einen Fold als Testdaten verwenden.
  - Die übrigen k−1 Folds als Trainingsdaten verwenden.
  - Modell trainieren und testen.
  - Vorgang k-mal wiederholen, sodass jeder Fold einmal Testdaten ist.
  - Die Ergebnisse mitteln.

Evaluation: Cross validation, Beispiel

Durchlauf	Training	Test
1	Fold 2–5	Fold 1
2	Fold 1, 3–5	Fold 2
3	Fold 1–2, 4–5	Fold 3
4	Fold 1–3, 5	Fold 4
5	Fold 1–4	Fold 5

Evaluation: Cross validation

- Rechenaufwand: Da das Modell mehrfach trainiert wird, benötigt Cross Validation mehr Rechenzeit als eine einfache Train-Test-Aufteilung.
- Wahl von k
  - Kleines k (z. B. 5): schneller
  - Grosses k (z. B. 10): stabilere Schätzung, aber höherer Aufwand
- Datenleck (Data Leakage): Informationen aus den Testdaten dürfen nicht in das Training gelangen, da sonst die Ergebnisse zu optimistisch werden.
- Zeitreihendaten: Normale k-Fold Cross Validation eignet sich nicht für Zeitreihen, da zukünftige Daten nicht zur Vorhersage vergangener Daten verwendet werden dürfen.

Evaluation: Cross validation

Vorteile	Nachteile
Zuverlässigere Leistungsbewertung	Höherer Rechenaufwand
Erkennt Overfitting	Kann bei grossen Datensätzen langsam sein
Nutzt Daten effizient	Nicht direkt für Zeitreihen geeignet
Gut für Modellvergleich	Sorgfältige Umsetzung erforderlich

Evaluation: Vergleich

Methode	Zweck	Vorteile	Nachteile	Typischer Einsatz
MAE (Mean Absolute Error)	Durchschnittliche absolute Fehler messen	Einfach verständlich, robust gegenüber Ausreissern	Grosse Fehler werden nicht besonders stark bestraft	Wenn die durchschnittliche Fehlergrösse wichtig ist
MSE (Mean Squared Error)	Durchschnittliche quadratische Fehler messen	Bestraft grosse Fehler stark	Stark empfindlich gegenüber Ausreissern, schwerer interpretierbar	Wenn grosse Fehler besonders vermieden werden sollen
RMSE (Root Mean Squared Error)	Fehler in Originaleinheit messen	Gleiche Einheit wie die Zielvariable, berücksichtigt grosse Fehler stärker	Sensibel gegenüber Ausreissern	Standardmetrik für viele Regressionsprobleme
Cross Validation	Generalisierungsfähigkeit bewerten	Erkennt Overfitting, liefert robustere Leistungsbewertung	Höherer Rechenaufwand	Modellvergleich und Modellvalidierung

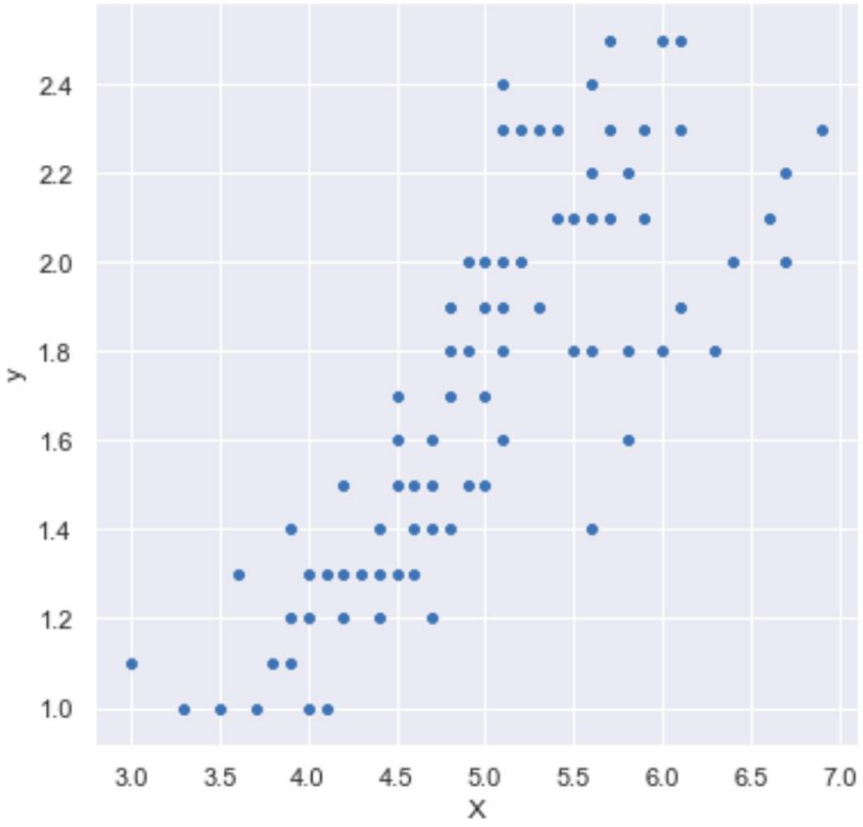
Evaluation: Herausforderungen

- Overfitting: Das Modell funktioniert sehr gut auf Trainingsdaten, aber schlecht auf neuen Daten.
- Underfitting: Das Modell ist zu einfach und erfasst wichtige Zusammenhänge nicht.
- Ausreisser: Einzelne extreme Werte können MSE und RMSE stark beeinflussen.
- Datenleck (Data Leakage): Informationen aus den Testdaten gelangen unbeabsichtigt in das Training und verfälschen die Evaluation.
- Unausgewogene Daten: Bestimmte Wertebereiche können über- oder unterrepräsentiert sein.

Regresion: Praxis

- zur Visualisierung von Ideen und Verfahren wird auch hier ein Demo Dataset verwendet:
  - demo_data_regr.csv
- zwei Spalten (columns)
  - X: Feature (unabhängige Variable)
  - y:Target (abhängige Variable)
- 81 Beobachtungen

- (es sind tatsächlich dieselben Daten wie bei demo_data_class.csv, ausser
  - X1 → X
  - X2 → y)



Regresion: Praxis

- wie bei der Klassifikation wird auch bei Regression mit einem Dataset aus einer konkreten Fallstudie gearbeitet
- das für die Praxisteile im Rahmen von Regression verwendete Dataset wurde im Rahmen des Workshops 03 unter Feature Engineering bereits aufbereitet
- einige Kennwerte
  - Anzahl rows: 18'393
  - Anzahl columns: 24,
    davon
    - float64: 10
    - int64:14
  - Target: "Price" (float64)
- Ziel der Arbeiten mit diesem Dataset: trainieren eines Vorhersagemodells für den Verkaufspreis von Immobilien

Regresion: Praxis

- wie bei den Methoden zur Klassifikation hat es auch bei der Regression im begleitenden Jupyter Notebook gleich am Anfang einen Codeblock, in welchem die Umgebung und die Daten vorbereitet werden:
  - importieren der notwendigen Libraries
  - setzen des Datenpfades
  - Laden und vorbereiten der Datasets
    - Demo Dataset
    - Melbourne Housing Dataset

- das Demo Dataset wird auch hier **nicht** in Train - Test gesplittet, es wird ausschliesslich dazu verwendet, die Regressionsmethoden darzustellen (X_demo, y_demo)
- die Performance Vergleiche erfolgen dann aber auf dem Melbourne Housing Dataset, welches aus diesem Grund gesplittet wird (X_train, y_train, X_test, y_test)

Regresion: Praxis

- die Standard Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

- analog dem Vorgehen bei Klassifikation können auch hier die im Modul bfh_cas_pml implementierten Funktionen für die Bereitstellung der Daten verwendet werden

```
from bfh_cas_pml import prep_data, prep_demo_data
X_train, X_test, y_train, y_test = prep_data(
    'melb_data_prep.csv', 'Price', seed = 1234)
X_demo, y_demo = prep_demo_data('demo_data_regr.csv', 'y')
```

Regresion: Praxis

- Schritt für Schritt mit dem Melbourne Housing Dataset (beachten Sie das analoge Vorgehen wie bei Klassifikation)
- Voraussetzungen:
  - Daten sind geladen
  - Features - Target - Split ausgeführt
  - Train - Test - Split ausgeführt

Regresion: Praxis

- laden der Klasse, instanziieren, parametrisieren und trainieren des Modells

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
print(model.get_params())
```

```
{'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'positive': False}
```

- `.get_params()` zeigt die per Default eingestellten Parameterwerte
- hier einzig von Bedeutung: fit_intercept, veranlasst, dass Intercept berechnet wird, andernfalls wird dieser Wert auf 0 festgelegt

Regresion: Praxis

- die Attribute des trainierten Modells:

```
print(model.intercept_)
print(model.coef_)
print(X_train.columns) ## no model attribute
```

```
-105513873.23401378
[ 2.45383606e+05 -1.41356398e+05 -4.03836664e+04  1.61336039e+05
  4.03911483e+04  8.33032709e+04  2.73783998e+05 -2.48422914e+03
:
Index(['Rooms', 'Type', 'Distance', 'Bathroom', 'Car', 'logLandsize',
       'logBuildingArea', 'YearBuilt', 'CouncilArea', 'Lattitude',
:
```

  - `X_train.columns` gibt die Feature Namen aus, damit die ermittelten Koeffizienten zugeordnet werden können

Regresion: Praxis

- Modell score

```
print(model.score(X_test, y_test))
```

```
0.5601419746121148
```

  - gemäss Dokumentation wird hier `r2_score` ermittelt

- Kontrolle mit predict und expliziter Nutzung von `r2_score` aus `sklearn.metrics`:

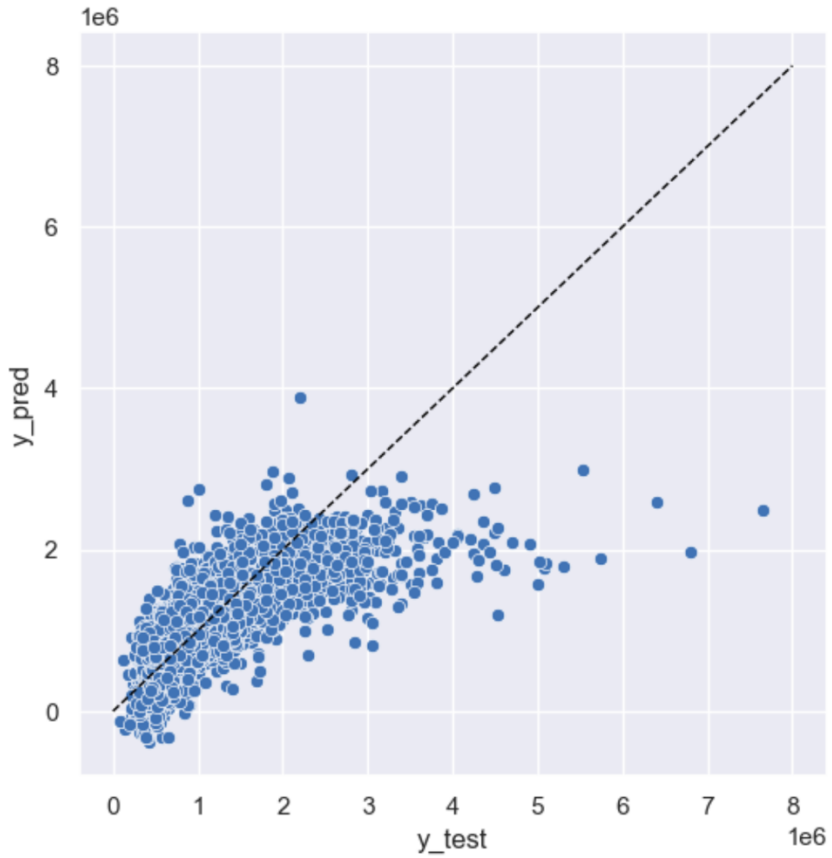
```
from sklearn.metrics import r2_score
y_pred = model.predict(X_test)
print(r2_score(y_test, y_pred))
```

```
0.5601419746121148
```

Regresion: Praxis

- ein visueller Vergleich der vorhergesagten (y_pred) und der wahren Targetwerte des Testsets (y_test)
- die eingezeichnete Diagonale wurde manuell hinzugefügt und markiert Identität

- Fazit:
  - dieses Modell erzeugt tatsächlich Vorhersagen mit negativen Werten, was doch recht unglaubwürdig erscheint
  - der Zusammenhang zwischen Voraussagen und wahren Werten erscheint nicht linear, ev. aufgrund von Korrelationen in den Features
  - welche Rolle spielen Extremwerte in den wahren Werten?



Regresion: Praxis

- Lasso

```
from sklearn.linear_model import Lasso
model = Lasso()
model.fit(X_train, y_train)
print(model.intercept_)
print(model.coef_)
print(model.score(X_test, y_test))
```

```
-105470886.17680839
[ 2.45382007e+05 -1.41353663e+05 -4.03813210e+04  1.61339539e+05
  4.03901128e+04  8.33016390e+04  2.73752966e+05 -2.48435395e+03
:
0.5601427046293168
```

Regresion: Praxis

- Ridge

```
from sklearn.linear_model import Ridge
model = Ridge()
model.fit(X_train, y_train)
print(model.intercept_)
print(model.coef_)
print(model.score(X_test, y_test))
```

```
-104848432.75507241
[ 2.45313734e+05 -1.41286745e+05 -4.03551367e+04  1.61451266e+05
  4.03883258e+04  8.32645429e+04  2.73073082e+05 -2.48815619e+03
:
0.560138763183778 4
```

- zur Erinnerung, der Parameter alpha entspricht λ in der theoretischen Einführung

Regresion: Praxis

Fazit zu Lasso und Ridge Regression

- als Kandidaten im Wettstreit der Regressoren nicht unbedingt Favorit
- **aber**: wichtige Instrumente für Feature Selection, insbesondere Lasso
- eine Darstellung der übrigbleibenden Feature Namen und Koeffizienten (nach trainieren mit dem Parameter `alpha=10000`) zeigt folgende Liste der verbleibenden Features an (vgl. [ipynb])
- von 23 Features bleiben also deren 7 übrig, aus dem Ergebnis könnte damit z.B. eine Filtermaske gebaut werden (vgl. [ipynb])

	cols	coefs
0	Rooms	253568.228957
3	Bathroom	168521.486901
4	Car	34293.196859
5	logLandsize	76942.997592
12	Method_S	18550.768445
17	Regionname_Southern_Metropolitan	286752.054486
22	day_of_week	781.020849

Evaluation: Workshop 08

Zeit: 60'

- untersuchen Sie den Einfluss des Standardisierens der Features auf folgende Ergebnisse der Linearen Regression:
  - Modellkoeffizienten
  - Predictions
  - Score

- *Wer gerne möchte oder zu Hause oder … : Untersuchen Sie den Einfluss des Logarithmierens des Targets auf die Performance der Linearen Regression*