

# QAF: Quantitative Forschungsmethoden

Tarek Carls

29. Oktober 2024

# Agenda

- **Session 1:** Grundlagen, induktive Statistik, Konfidenzintervalle
- **Session 2:** t-Tests, einfaktorielle ANOVA
- **Session 3:** Mehrfaktorielle ANOVA
- **Session 4:** Lineare Regression, logistische Regression
- **Session 5:** Fragen und Wiederholung

# Agenda - Session 3

- 1 Korrelation
- 2 Lineare Regression
- 3 Logistische Regression

# Agenda - Session 3

- 1 Korrelation
- 2 Lineare Regression
- 3 Logistische Regression

# Einführung in die Korrelation

Korrelation ist ein statistisches Maß, das die Stärke und Richtung einer linearen Beziehung zwischen zwei Variablen misst.

- Zeigt, wie sich eine Variable ändert, wenn eine andere variiert
- Wichtig für die Vorhersage und Modellierung in vielen Forschungsfeldern
- Nicht gleichzusetzen mit Kausalität

# Der Korrelationskoeffizient

Der Korrelationskoeffizient (üblicherweise Pearson's  $r$ ) quantifiziert die Korrelation.

- Wertebereich: -1 bis 1
- Interpretation der Werte:
  - Nahe 0: Schwache Korrelation
  - Nahe +1: Starke positive Korrelation
  - Nahe -1: Starke negative Korrelation

Formel für Pearson's  $r$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

# Typen von Korrelation

- Pearson-Korrelation: Misst lineare Beziehungen zwischen metrischen Variablen.
- Spearman-Korrelation: Eignet sich für rangbasierte Daten oder nicht-lineare Beziehungen.
- Kendall-Tau: Eine weitere Rangkorrelation für ordinalskalierte Daten.
- Grafische Darstellung über Streudiagramme: Visualisieren die Beziehung zwischen zwei Variablen.
- Korrelationsmatrix: Nützlich bei der Untersuchung mehrerer Variablen.

# Agenda - Session 3

- 1 Korrelation
- 2 Lineare Regression
- 3 Logistische Regression



# Einführung in die lineare Regression

Lineare Regression ist eine statistische Methode zur Modellierung der Beziehung zwischen einer abhängigen Variable und einer oder mehreren unabhängigen Variablen.

- Ziel: Vorhersage oder Erklärung der abhängigen Variable (Y) basierend auf den unabhängigen Variablen (X).

# Das lineare Regressionsmodell

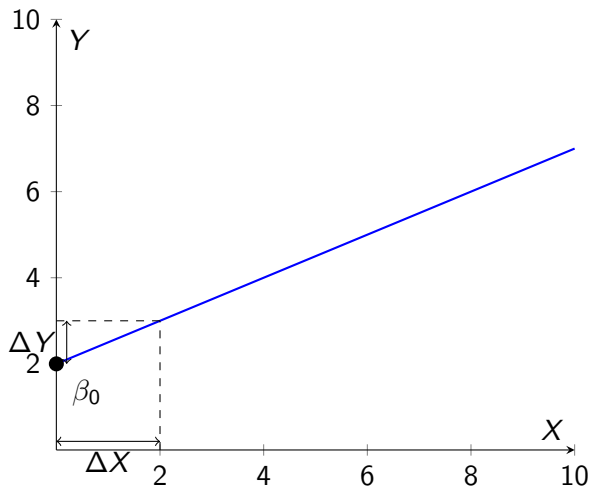
Das grundlegende lineare Regressionsmodell für eine abhängige Variable und eine unabhängige Variable ist:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $\beta_0$  (Achsenabschnitt) und  $\beta_1$  (Steigung): Zu schätzende Parameter.
- $\epsilon$ : Zufallsfehlerterm, der die Abweichung der Datenpunkte von der Regressionslinie darstellt.

# Lineare Beziehung im Koordinatensystem

$$Y = \beta_0 + \beta_1 X$$



# Bestimmung der Regressionskoeffizienten

Die Koeffizienten  $\beta_0$  und  $\beta_1$  werden durch die Methode der kleinsten Quadrate (OLS) bestimmt.

- OLS minimiert die Summe der quadrierten Differenzen zwischen beobachteten und vorhergesagten Werten.
- Mathematische Optimierung führt zu den besten Schätzern für  $\beta_0$  und  $\beta_1$ .

# Interpretation der Koeffizienten

- $\beta_0$ : Wert von Y, wenn alle X gleich null sind.
- $\beta_1$ : Durchschnittliche Änderung in Y für eine Einheitsänderung in X.
- Wichtig: Kausalinterpretation nur zulässig, wenn bestimmte Bedingungen erfüllt sind (z.B. keine Verzerrung durch ausgelassene Variablen).

# Voraussetzungen für lineare Regression

Wichtige Annahmen für die Anwendung der linearen Regression:

- Linearität: Die Beziehung zwischen  $X$  und  $Y$  ist linear.
- Unabhängigkeit: Beobachtungen sind unabhängig voneinander.
- Homoskedastizität: Konstante Varianz der Fehler über alle Werte von  $X$ .
- Normalverteilung der Fehler: Die Residuen (Fehler) sind normalverteilt.
- Keine perfekte Multikollinearität: Bei mehreren unabhängigen Variablen sollten diese nicht perfekt korrelieren.

# Güte des Modells

Die Güte des Modells wird durch mehrere Statistiken bewertet:

- Bestimmtheitsmaß ( $R^2$ ): Anteil der Varianz von  $Y$ , der durch das Modell erklärt wird.
- Adjustiertes  $R^2$ : Berücksichtigt die Anzahl der Prädiktoren im Modell.
- F-Test: Prüft, ob das Modell insgesamt signifikant ist.
- p-Werte der Koeffizienten: Testen die Hypothese, dass einzelne Koeffizienten gleich null sind.

# Agenda - Session 3

- 1 Korrelation
- 2 Lineare Regression
- 3 Logistische Regression



# Einführung in die logistische Regression

Die logistische Regression wird zur Vorhersage der Wahrscheinlichkeit eines binären Ergebnisses genutzt.

- Geeignet für abhängige Variablen mit zwei Kategorien (z.B. Ja/Nein, Erfolg/Misserfolg).
- Im Gegensatz zur linearen Regression, wo die abhängige Variable kontinuierlich ist.
- Anwendungen umfassen Medizin, Finanzen, Marketing und mehr.

# Das logistische Regressionsmodell

Die logistische Funktion (auch Logit-Funktion genannt) wird verwendet, um die Wahrscheinlichkeit zu modellieren.

$$\log \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

- Link-Funktion: Verbindet die lineare Kombination der Prädiktoren mit der Wahrscheinlichkeit des Ergebnisses.
- Die rechte Seite der Gleichung ist ein lineares Modell.

# Interpretation der Koeffizienten

Koeffizienten in der logistischen Regression haben eine spezifische Interpretation.

- Änderungen der unabhängigen Variablen beeinflussen die Odds (Chancen) des Ereignisses.
- Ein positiver Koeffizient erhöht die Odds, ein negativer verringert sie.
- Odds Ratio: Exponent der Koeffizienten zeigt die Veränderung der Odds für eine Einheitsänderung der unabhängigen Variable.

# Voraussetzungen und Herausforderungen

Einige wichtige Aspekte und Herausforderungen bei der logistischen Regression:

- Keine Multikollinearität: Unabhängige Variablen sollten nicht hochkorreliert sein.
- Große Stichprobengrößen sind oft erforderlich, um genaue Schätzungen zu erhalten.
- Überanpassung (Overfitting) und Unteranpassung (Underfitting) des Modells müssen berücksichtigt werden.
- Auswahl relevanter Variablen und Interaktionen ist entscheidend.

# Modellbewertung und Validierung

Methoden zur Bewertung der Güte und Validierung des logistischen Regressionsmodells:

- Konfusionsmatrix: Gibt Einblick in die Leistung des Modells (z.B. Sensitivität, Spezifität).
- Pseudo- $R^2$ : Bietet eine Einschätzung der Erklärungskraft des Modells.
- AUC-ROC-Kurve: Bewertet die Leistungsfähigkeit des Modells über verschiedene Klassifizierungsschwellen.
- Kreuzvalidierung: Beurteilt die Generalisierbarkeit des Modells auf neue Daten.