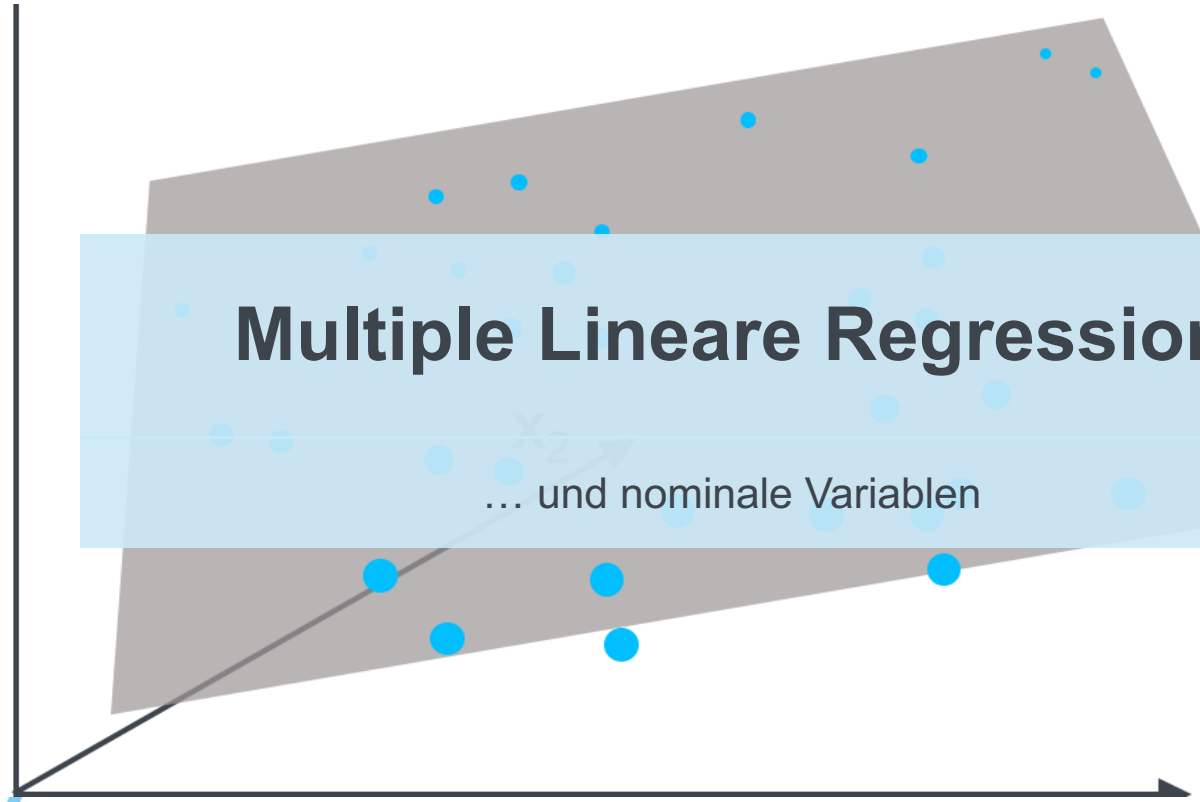


# Multiple Lineare Regression

… und nominale Variablen

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

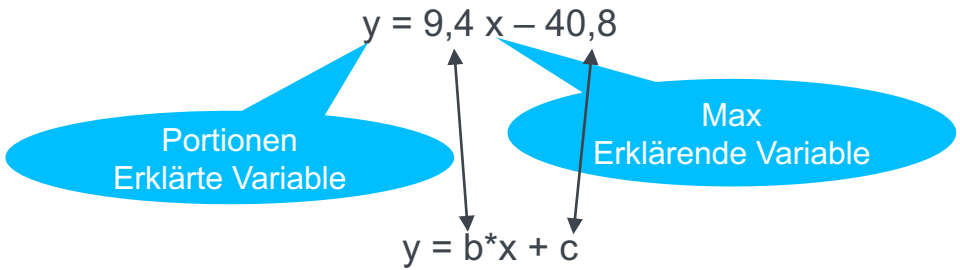
INNOVET
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

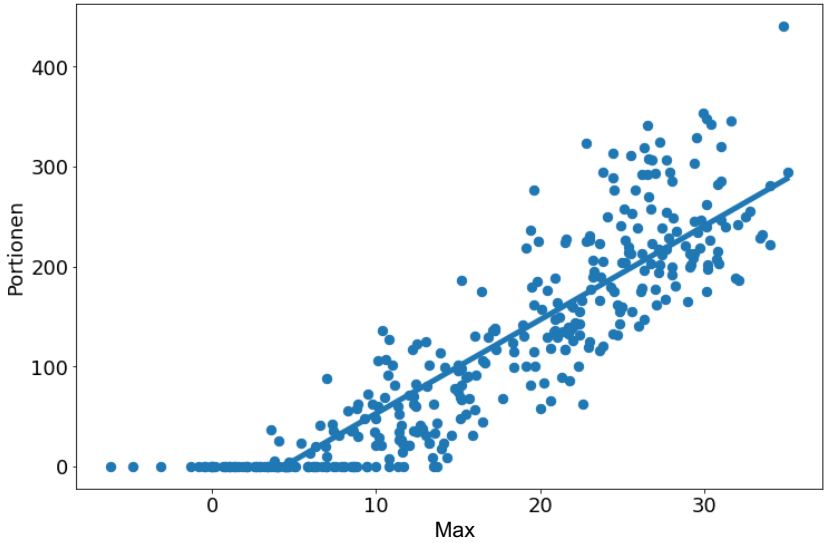
Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

# Rückblick:
# einfache lineare Regression

# Einfache lineare Regression

- Regressionslinie:




- Hier b = 9,4 ⇒ jede Erhöhung von x um eins bewirkt eine Erhöhung von y um 9,4
- Hier c = -40,8 ⇒ dieser Wert wird erwartet, wenn x = 0 ist (hier nicht sinnvoll zu interpretieren)

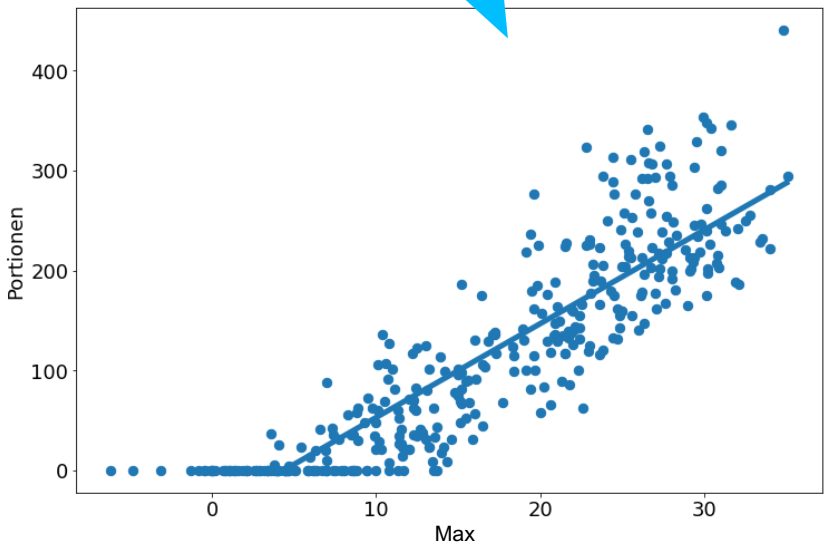
# Multiple lineare Regression

- Was, wenn es weitere erklärende Variablen gibt?

$$y = b*x + c$$

- Wir ergänzen einfach:
  - $y = b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n + c$
- Unverändert: Finden der Parameter $b_1$ bis $b_n$ und c durch Minimierung der Fehlerquadrate
- Visualisierung: schwierig 😬

Visualisierung im zweidimensionalen Fall: eine erklärende und eine erklärte Variable

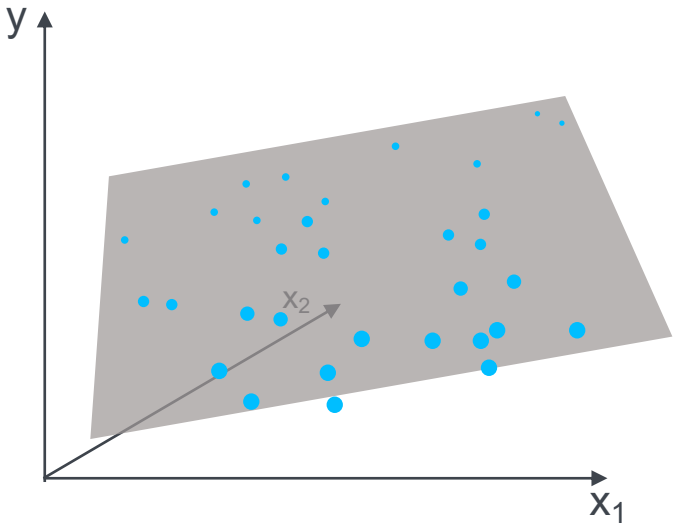


# Visualisierung im dreidimensionalen Fall

- Drei Dimensionen:
  - 2 unabhängige (erklärende) Variablen
  - 1 abhängige (erklärte Variable)

$$y = b_1 * x_1 + b_2 * x_2 + c$$

  - Zu modellierende Daten als dreidimensionale Wolke im Raum
  - Statt Annäherung durch Regressionsgerade: Annäherung durch Regressionsebene



# Parameter bei multipler linearer Regression

- Einfache lineare Regression:
  - $y = b*x + c$
  - Steigung b und y-Achsenabschnitt c
- Multiple lineare Regression:
  - $y = b_1*x_1 + b_2*x_2 + b_3*x_3 + \ldots + b_n*x_n + c$
  - Viele „Steigungen“ $b_1$ bis $b_n$
  - Konzept der Steigung im mehrdimensionalen Raum nicht mehr so eingängig
    ⇒ man spricht meist nur noch von „Koeffizienten“ $b_1$ bis $b_n$
  - Manchmal wird auch der y-Achsenabschnitt nicht extra bezeichnet, z.B.:
    $y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + \ldots + b_n*x_n$

1 erkl. Variable
2 Parameter

n erkl. Variablen
n+1 Parameter

**Bei multipler linearer Regression bekommt man wie bei der einfachen linearen Regression einen Parameter für den y-Achsenabschnitt.**

**Außerdem für jede unabhängige Variable einen Koeffizienten, der den Einfluss dieser Variable auf die abhängige Variable beschreibt.**

# Prognosen mit solchen Modellen

# Vorhersage von Immobilienwerten („California Housing“ Daten)

| | Wert | Einkommen | Hausalter | Zimmer/Haushalt | Schlafzimmer/Haushalt | Einwohner | Bewohner/Haushalt | West | Nord |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 4.526 | 8.3252 | 41.0 | 6.984127 | 1.023810 | 322.0 | 2.555556 | 37.88 | -122.23 |
| **1** | 3.585 | 8.3014 | 21.0 | 6.238137 | 0.971880 | 2401.0 | 2.109842 | 37.86 | -122.22 |
| **2** | 3.521 | 7.2574 | 52.0 | 8.288136 | 1.073446 | 496.0 | 2.802260 | 37.85 | -122.24 |
| **3** | 3.413 | 5.6431 | 52.0 | 5.817352 | 1.073059 | 558.0 | 2.547945 | 37.85 | -122.25 |
| **4** | 3.422 | 3.8462 | 52.0 | 6.281853 | 1.081081 | 565.0 | 2.181467 | 37.85 | -122.25 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **20635** | 0.781 | 1.5603 | 25.0 | 5.045455 | 1.133333 | 845.0 | 2.560606 | 39.48 | -121.09 |
| **20636** | 0.771 | 2.5568 | 18.0 | 6.114035 | 1.315789 | 356.0 | 3.122807 | 39.49 | -121.21 |
| **20637** | 0.923 | 1.7000 | 17.0 | 5.205543 | 1.120092 | 1007.0 | 2.325635 | 39.43 | -121.22 |
| **20638** | 0.847 | 1.8672 | 18.0 | 5.329513 | 1.171920 | 741.0 | 2.123209 | 39.43 | -121.32 |
| **20639** | 0.894 | 2.3886 | 16.0 | 5.254717 | 1.162264 | 1387.0 | 2.616981 | 39.37 | -121.24 |

abh. V. | unabhängige Variablen

# Berechnung für ersten Datenpunkt

| | Einkommen | Hausalter | Zimmer/Haushalt | Schlafzimmer/Haushalt | Einwohner | Bewohner/Haushalt | West | Nord |
|---|---|---|---|---|---|---|---|---|
| 0 | 8.3252 | 41.0 | 6.984127 | 1.023810 | 322.0 | 2.555556 | 37.88 | -122.23 |

| | Variable | Koeffizient |
|---|---|---|
| 0 | Einkommen | 0.436693 |
| 1 | Hausalter | 0.009436 |
| 2 | Zimmer/Haushalt | -0.107322 |
| 3 | Schlafzimmer/Haushalt | 0.645066 |
| 4 | Einwohner | -0.000004 |
| 5 | Bewohner/Haushalt | -0.003787 |
| 6 | West | -0.421314 |
| 7 | Ost | -0.434514 |
| y-Achsenabschnitt: | | -36.941920 |

Wert = 0.436693 * 8.3252
+ 0.009436 * 41
- 0.107322 * 6.984127
+ 0.645066 * 1.023810
- 0.000004 * 322
- 0.003787 * 2.555556
- 0.421314 * 37.88
-0.434514 * (-122.23)
-36.94192021
= 4.13169

nicht so schlecht:
korrekt wäre
4.526

# Quiz - Multiple Choice: Abhängige/Unabhängige Variablen

# Nominale Variablen

# Beispiel Eisdaten

- Erklärende Variablen bisher: Temperatur, Bewölkung
  - Beide sind „messbare Größen“
  - Ihre Werte können mit einem Koeffizienten multipliziert werden
- Wochentag als weitere Einflussgröße: an Wochenenden mehr Verkäufe
  - Keine messbare Größe, nicht einmal „ordinal skaliert“, d.h. nicht nach Größe sortierbar (Montag $\prec$ Dienstag $\prec$ Mittwoch $\prec$ … $\prec$ Sonntag)

Wochentag: kategoriale (nominale) Variable

Temperatur als unabh. Variable: messbare Größe

Koeff.: b

20°C < 30°C
wenn b positiv:
b * 20 < b * 30

Was tun bei kategorialen Variablen?

**Kategoriale Variablen (auch: nominale Variablen) sind Variablen, die abzählbar viele, nicht nach der Größe sortierbare mögliche Werte haben.**

# Binäre Variablen

- Variablen mit 2 möglichen Werten
- Wochenende ja/nein kann als 1 bzw. 0 kodiert werden
- Bei nur zwei Werten ist die Ordnung kein Problem:
  - Angenommen, an Wochenenden wird mehr Eis verkauft
  - 2 mögliche Kodierungen
  - Koeffizient zeigt den Effekt des Merkmals, das als 1 kodiert ist
- Funktioniert bei allen kategorialen Variablen, die nur zwei Kategorien haben

Wochenende ja: 1
Wochenende nein: 0
⇒ positiver Koeffizient b

Bsp.: b = 5

$5 * 1 > 5 * 0$

$5 > 0$

Wochenende ja: 0
Wochenende nein: 1
⇒ negativer Koeffizient b

Bsp.: b = -5

$-5 * 0 > -5 * 1$

$0 > -5$

# Kategoriale Variablen mit mehreren Kategorien

Ist_Sonntag: redundant: Sonntag durch 0 in allen Spalten erfasst

| Wochentag |
| --- |
| Montag |
| Dienstag |
| Mittwoch |
| Donnerstag |
| Freitag |
| Samstag |
| Sonntag |

- Mehrere **Dummy-Variablen** zur Kodierung

**Kategoriale Variablen können mithilfe von Dummy-Kodierung als unabhängige Variablen in Regressionsmodellen verwendet werden.**

**Bei der Dummy-Kodierung vergibt man den Wert 1, wenn es sich um eine bestimmte Kategorie handelt, sonst 0.**

**Bei einer Variable mit n Kategorien benötigt man n-1 Dummy-Variablen.**

# Quiz - Multiple Choice: Dummy-Kodierung

# Quiz - Drag and Drop: Nominale Variablen

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen


**Dr. Antje Schweitzer**

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung


Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München


GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung





Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

# Lizenzbestimmungen

„Multiple Lineare Regression“ von Antje Schweitzer, KI B$^3$ / Uni Stuttgart



**Quellenverzeichnis**

Seite 12, Clipart: Microsoft