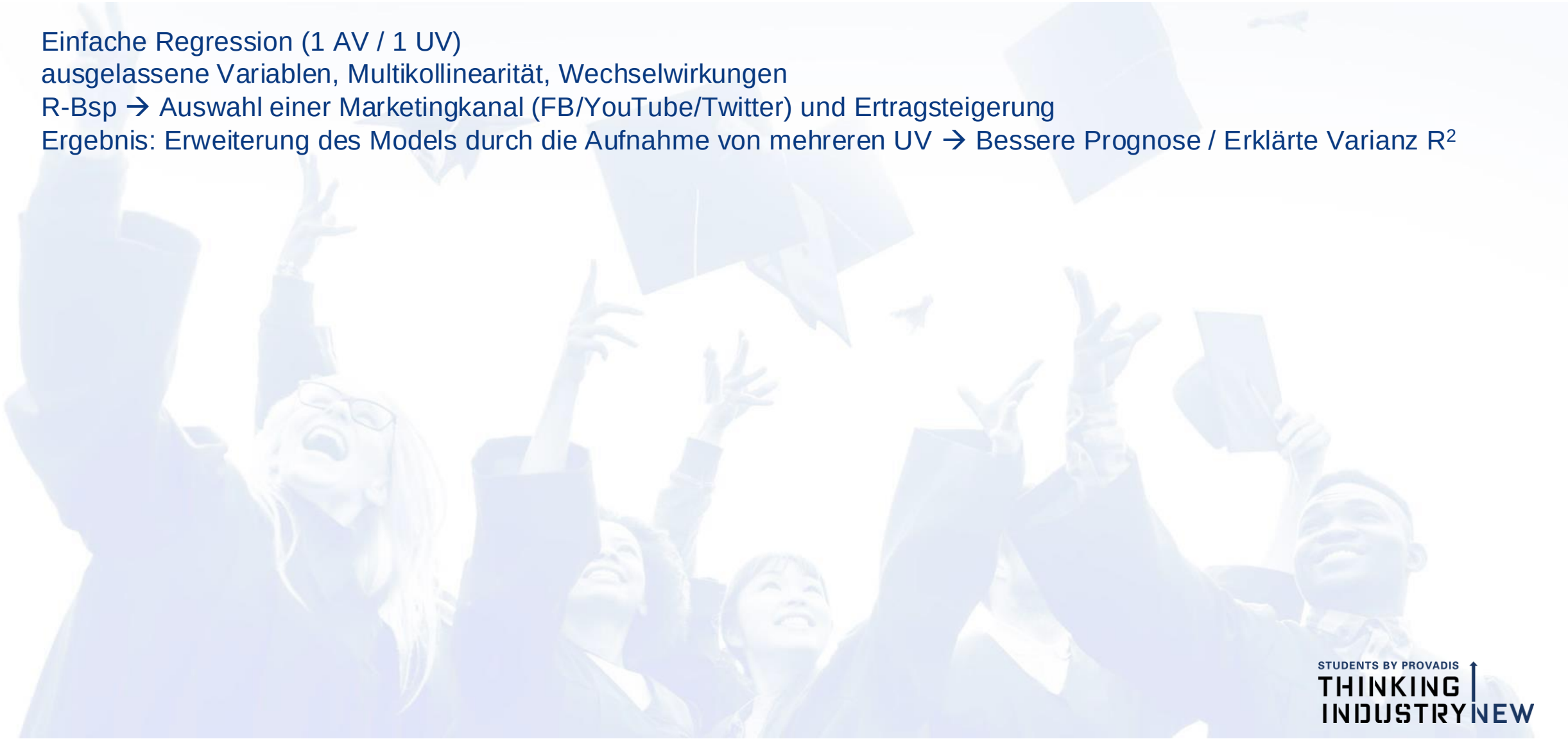
Multiple Regression

Dr. Houssam Jedidi

Wo wir waren und wo wir hinwollen...

provadis Hochschule

Einfache Regression (1 AV / 1 UV)
ausgelassene Variablen, Multikollinearität, Wechselwirkungen
R-Bsp → Auswahl einer Marketingkanal (FB/YouTube/Twitter) und Ertragsteigerung
Ergebnis: Erweiterung des Models durch die Aufnahme von mehreren UV → Bessere Prognose / Erklärte Varianz R^2



STUDENTS BY PROVADIS
THINKING INDUSTRY NEW

provadis Hochschule

Grundlagen

- **Abhängige und unabhängige Variablen:**
  - Abhängige Variable (y): Die Variable, die erklärt oder vorhergesagt werden soll.
  - Unabhängige Variable (x): Die Variable, die zur Erklärung oder Vorhersage verwendet wird.

- **Lineares Modell:** $y = \beta_0 + \beta_1 x + \varepsilon$

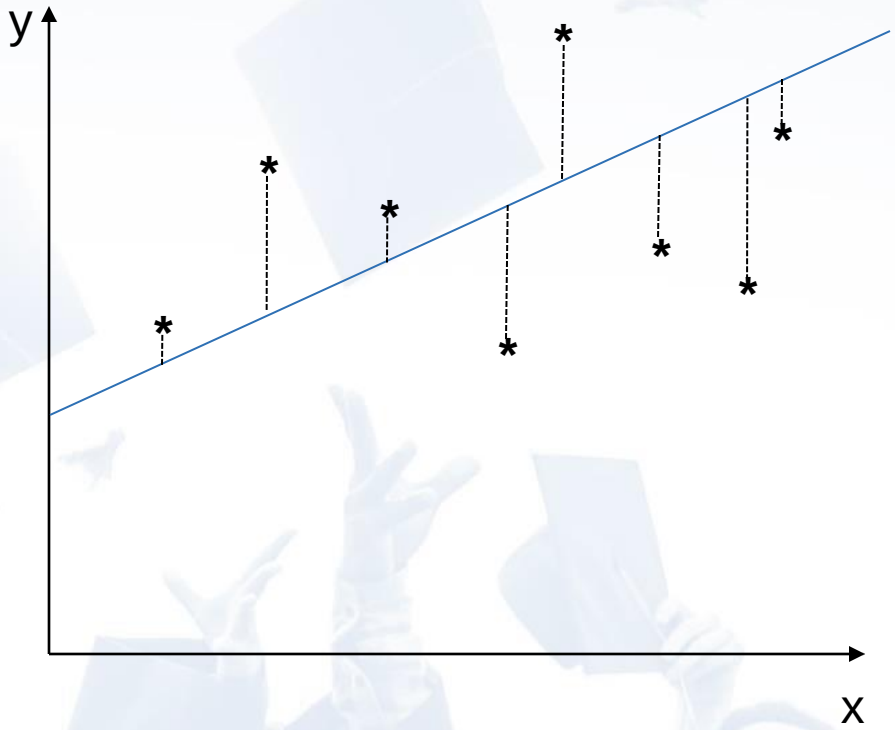
/ $ß_0$: $Achsenabschnitt$
$ß_1$: $Regressionskoeffizient\ (Steigung)$
ε: $Fehlerterm$

Zweidimensionale Verteilungen: Korrelations– und Regressionsanalyse
Regressionsanalyse

- **Graphische Darstellung:**
  - Streudiagramm mit Regressionsgerade
  - $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$

- **Berechnung der Koeffizienten:**
  - $ß_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

  - $ß_0 = \bar{y} - ß_1 \bar{x}$



provadis Hochschule

Interpretation der Regressionskoeffizienten:

- **Achsenabschnitt (β_0):** Wert von y, wenn x = 0.
- **Steigung (β_1):** Veränderung von y, wenn x um eine Einheit zunimmt.
- **Beispiel:** Wenn $\beta_0=2$ und $\beta_1=0.5$, dann bedeutet dies, dass y um 0.5 Einheiten steigt, wenn x um eine Einheit zunimmt.

Wichtig: Die Interpretation der Koeffizienten muss im Kontext der Daten erfolgen.

provadis Hochschule

Modellbewertung und Gütekriterien

Bestimmtheitsmaß (R^2): Maß für die Güte der Anpassung

$$R^2 = \frac{Erklärte\ Varianz}{Gesmatvarianz}$$

Wertebereich zw. 0 und 1 (je näher an 1, desto besser die Anpassung)

Standardfehler der Schätzung: Maß für die durchschnittliche Abweichung der beobachteten Werte von der Regressionsgeraden.

F-Statistik: Testet die Gesamtbedeutung des Modells.

Zweidimensionale Verteilungen: Korrelations– und Regressionsanalyse
Regressionsanalyse

provadis Hochschule

•**Beispiel:** Untersuchung des Zusammenhangs zwischen Werbeausgaben (x) und Verkaufszahlen (y).

•**Datensatz:**
•x (Werbeausgaben in Tausend Euro): [1, 2, 3, 4, 5]
•y (Verkaufszahlen in Tausend Einheiten): [2, 4, 5, 4, 5]

- $ß_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = 0.7$
- $ß_0 = \bar{y} - ß_1 \bar{x} = 2{,}6$
- $\boldsymbol{y = 2,6 + 0,7.x}$

Multivariate Rregressionsmodelle

provadis Hochschule

STUDENTS BY PROVADIS
THINKING INDUSTRY NEW

Warum Matrizen & Vektoren?

Regressionsgleichung kann wie folgt beschrieben werden:

Die Notation wird unnötig unübersichtlich, wenn wir Variablen hinzufügen Matrizen sind sauber, aber sie sind wie eine Fremdsprache

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots . \beta_k x_{ik} + u_i$$

Sie müssen über einen langen Zeitraum hinweg Intuitionen entwickeln

Zur Erinnerung an die Interpretation der Parameter:
β1 ist die Auswirkung einer Änderung von x_{i1} um eine Einheit unter der Bedingung aller anderen x_{ik}.

Das lineare Modell mit neuer Notation

Erinnern Sie sich, dass wir das lineare Modell wie folgt für alle i ∈ [1,...,n] geschrieben haben:

Stellen Sie sich vor, wir hätten ein n von 4.

Wir könnten jede Formel aufschreiben:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$$

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + u_1 \quad \text{Einheit\_1}$$

$$y_2 = \beta_0 + \beta_1 x_2 + \beta_2 z_2 + u_2 \quad \text{Einheit\_2}$$

$$y_3 = \beta_0 + \beta_1 x_3 + \beta_2 z_3 + u_3 \quad \text{Einheit\_3}$$

$$y_4 = \beta_0 + \beta_1 x_4 + \beta_2 z_4 + u_4 \quad \text{Einheit\_4}$$

Das lineare Modell mit neuer Notation

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + u_1 \quad \text{Einheit\_1}$$
$$y_2 = \beta_0 + \beta_1 x_2 + \beta_2 z_2 + u_2 \quad \text{Einheit\_2}$$
$$y_3 = \beta_0 + \beta_1 x_3 + \beta_2 z_3 + u_3 \quad \text{Einheit\_3}$$
$$y_4 = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_4 \quad \text{Einheit\_4}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

Alles in Matrizen gruppieren

Wir können alle Koeffizienten in einer Matrix

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ 1 & x_4 & z_4 \end{bmatrix} \qquad \underset{(3\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Matrixmultiplikation mit einem Vektor

Wir können dies kompakter als eine mit einem Vektor multiplizierte Matrix schreiben:

Die Multiplikation einer Matrix mit einem Vektor ist einfach die Linearkombination der Spalten der Matrix mit den Vektorelementen als Gewichte/Koeffizienten.
Und die linke Seite verwendet hier nur Skalare mal Vektoren, was einfach ist!

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 = \mathbf{X}\boldsymbol{\beta}$$

provadis Hochschule

$$\begin{pmatrix} 2 & -5 \\ 13 & 7 \\ -6 & 4 \end{pmatrix} + \begin{pmatrix} 7 & 10 \\ -8 & 1 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} 2+7 & -5+10 \\ 13+(-8) & 7+1 \\ -6+0 & 4+(-3) \end{pmatrix} = \begin{pmatrix} 9 & 5 \\ 5 & 8 \\ -6 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & -5 \\ 13 & 7 \\ -6 & 4 \end{pmatrix} - \begin{pmatrix} 7 & 10 \\ -8 & 1 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} -5 & -15 \\ 21 & 6 \\ -6 & 7 \end{pmatrix}$$

provadis Hochschule

$$\begin{pmatrix} 2 & 5 \\ 6 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \cdot 3 + 5 \cdot 4 \\ 6 \cdot 3 + 1 \cdot 4 \end{pmatrix} = \begin{pmatrix} 26 \\ 22 \end{pmatrix}$$

$$\begin{pmatrix} 9 & 0 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 9 \cdot 5 + 0 \cdot 7 \\ 2 \cdot 5 + 1 \cdot 7 \end{pmatrix} = \begin{pmatrix} 45 \\ 17 \end{pmatrix}$$

Zurück zur Regression

X ist die n × (K + 1) Designmatrix der unabhängigen Variablen **β** ist der (K + 1) × 1 Spaltenvektor der Koeffizienten.
Xβ wird n×1 sein:

Wir können das lineare Modell kompakt wie folgt schreiben:

Wir können dies auch auf individueller Ebene schreiben, wobei **x′**i die i-te Zeile von **X** ist

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_K\mathbf{x}_K$$

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}}$$

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

provadis Hochschule

Hypothesenformulierung & Testing

STUDENTS BY PROVADIS
THINKING INDUSTRY NEW

Beispiel: Chilenisches Referendum über Pinochet

Das chilenische Plebiszit von 1988 war ein nationales Referendum, bei dem es darum ging, ob Diktator **Augusto Pinochet** seine Amtszeit um weitere acht Jahre verlängern würde oder nicht.
Daten: Nationale Umfrage, durchgeführt im April und Mai 1988 von FLACSO in Chile.

Ergebnis: % Wahlwahrscheinlichkeit für Pinochet (0→100%).

Wir können die β-Steigungen als marginale "Auswirkungen" auf die Wahrscheinlichkeit interpretieren, dass der Befragte für Pinochet stimmt.

Das Plebiszit wurde am 5. Oktober 1988 abgehalten. Die Nein-Seite gewann mit 56% der Stimmen, während 44% mit Ja stimmten (umkodieren).

Wir modellieren die beabsichtigte Pinochet-Stimme als eine lineare Funktion von Geschlecht, Bildung und Alter der Befragten.

Beispiel: Chilenisches Referendum über Pinochet

provadis Hochschule

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4875 on 1699 degrees of freedom
Multiple R-squared: 0.05112,        Adjusted R-squared: 0.04945
F-statistic: 30.51 on 3 and 1699 DF,  p-value: < 2.2e-16
```

Beispiel: Chilenisches Referendum über Pinochet

Modellinterpretation
Vorschläge & Lösungen

Signifikanzlevel (z. B. 0,05 oder 0,01) werden verwendet, um zu entscheiden, ob ein Ergebnis statistisch signifikant ist. Sie zeigen die Wahrscheinlichkeit, dass ein beobachtetes Ergebnis allein durch Zufall entsteht, wenn die Nullhypothese (H_0) wahr ist.

- **Nullhypothese (H_0):** Der Prädiktor hat keinen Einfluss auf die Zielvariable.
- **Alternativhypothese (H_1):** Der Prädiktor hat einen Einfluss auf die Zielvariable.

Beispiel: Chilenisches Referendum über Pinochet

Modellinterpretation
Vorschläge & Lösungen

Typische Signifikanzlevel

- **0,05 (5%)**: Es besteht eine 5%ige Wahrscheinlichkeit, dass der Effekt zufällig ist. Dies ist der Standardwert in den meisten Studien.
- **0,01 (1%)**: Strengeres Niveau. Weniger wahrscheinlich, dass der Effekt zufällig ist.
- **0,001 (0,1%)**: Sehr streng. Wird häufig in hochpräzisen Studien verwendet.

Ursprung der Signifikanzlevel

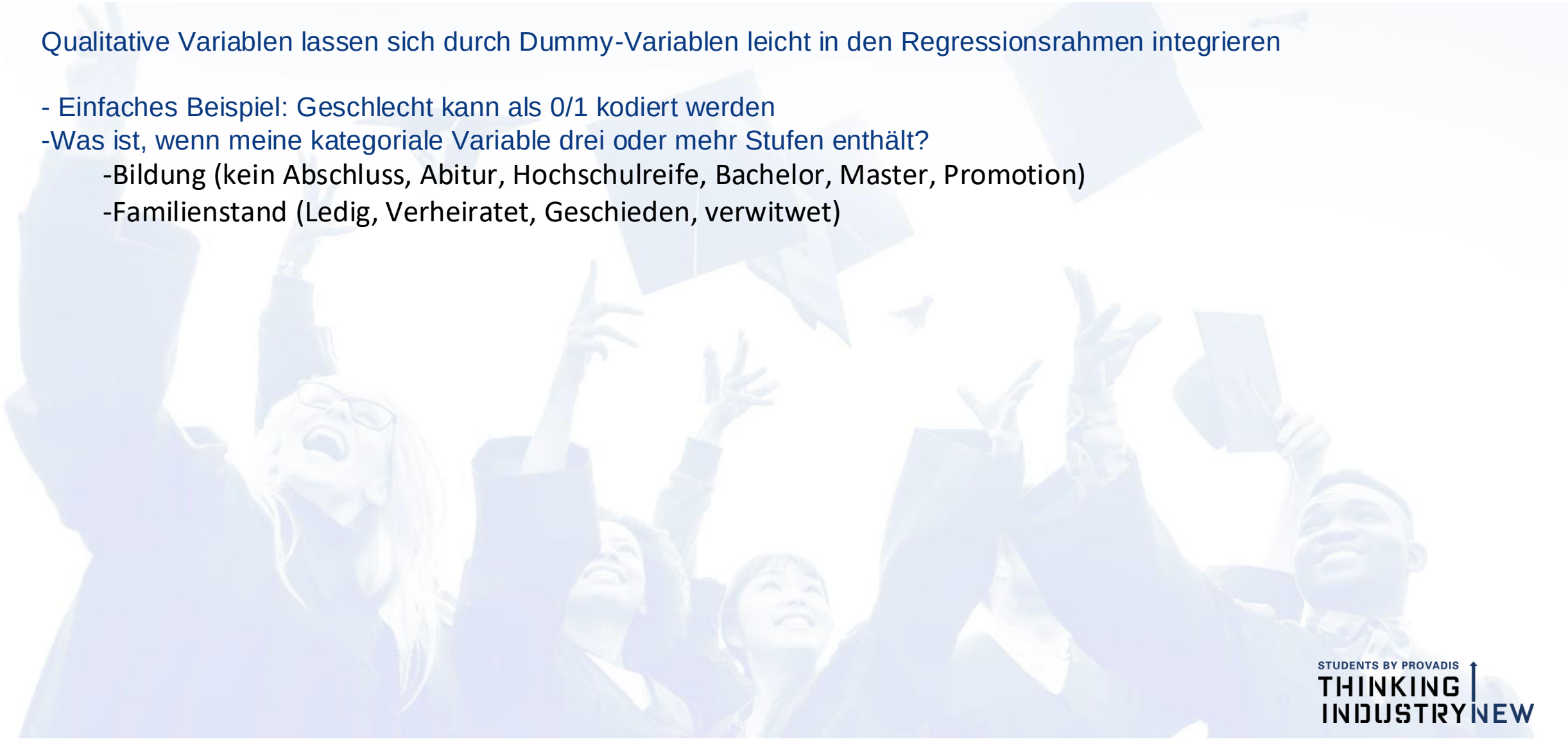
Die Idee stammt aus der Statistik und basiert auf dem Konzept von Hypothesentests. Das Signifikanzniveau wird oft als "Alpha" bezeichnet und legt den Schwellenwert fest, ab dem ein Ergebnis als signifikant angesehen wird.

Signifikanzlevel (z. B. 0,05 oder 0,01) werden verwendet, um zu entscheiden, ob ein Ergebnis statistisch signifikant ist. Sie zeigen die Wahrscheinlichkeit, dass ein beobachtetes Ergebnis allein durch Zufall entsteht, wenn die Nullhypothese (H_0) wahr ist.

Kategorische Variablen

Qualitative Variablen lassen sich durch Dummy-Variablen leicht in den Regressionsrahmen integrieren

- Einfaches Beispiel: Geschlecht kann als 0/1 kodiert werden
-Was ist, wenn meine kategoriale Variable drei oder mehr Stufen enthält?
    -Bildung (kein Abschluss, Abitur, Hochschulreife, Bachelor, Master, Promotion)
    -Familienstand (Ledig, Verheiratet, Geschieden, verwitwet)



Kategorische Variablen:
Kodierung hängt von den Hypothesen und Zielsetzung ab.

provadis Hochschule

id	Geschlecht
1	M
2	W
3	M
4	D

id	M	W	D
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1

id	Bildung
1	Abitur
2	Promotion
3	Master
4	Promotion

id	Bildung
1	2
2	5
3	4
4	5

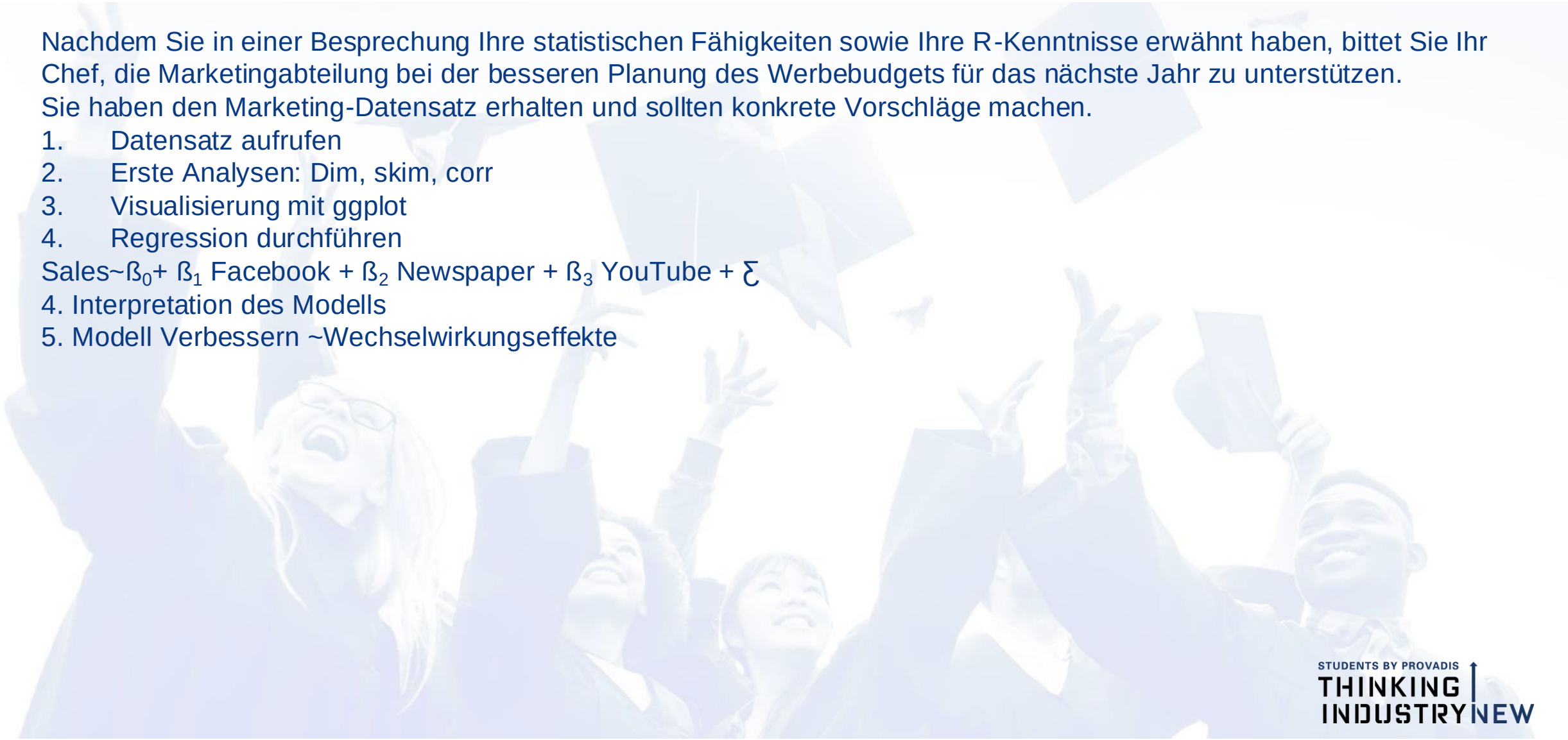
BSP: Marketing_Datensatz

Nachdem Sie in einer Besprechung Ihre statistischen Fähigkeiten sowie Ihre R-Kenntnisse erwähnt haben, bittet Sie Ihr Chef, die Marketingabteilung bei der besseren Planung des Werbebudgets für das nächste Jahr zu unterstützen.
Sie haben den Marketing-Datensatz erhalten und sollten konkrete Vorschläge machen.

1. Datensatz aufrufen
2. Erste Analysen: Dim, skim, corr
3. Visualisierung mit ggplot
4. Regression durchführen

Sales~$ß_0$+ $ß_1$ Facebook + $ß_2$ Newspaper + $ß_3$ YouTube + ξ

4. Interpretation des Modells
5. Modell Verbessern ~Wechselwirkungseffekte



Mtcars- Bsp.

Aufgaben lösen

Ziel
Analysieren Sie den Datensatz mtcars, um herauszufinden, welche Faktoren den Kraftstoffverbrauch (mpg) signifikant beeinflussen. Stellen Sie sicher, dass Sie die Modellannahmen überprüfen, die Ergebnisse interpretieren und eine klare Schlussfolgerung ziehen.

Explorative Datenanalyse (EDA):
•Laden Sie den Datensatz mtcars und verschaffen Sie sich einen Überblick über die Daten.
•Untersuchen Sie die Verteilung der Variablen und ihre Korrelationen.

Modellerstellung:
•Erstellen Sie ein multivariates lineares Regressionsmodell, um mpg basierend auf wt (Gewicht), hp (PS) und cyl (Zylinder) vorherzusagen.
•Geben Sie die Modellzusammenfassung (summary()) aus.

Mtcars- Bsp.

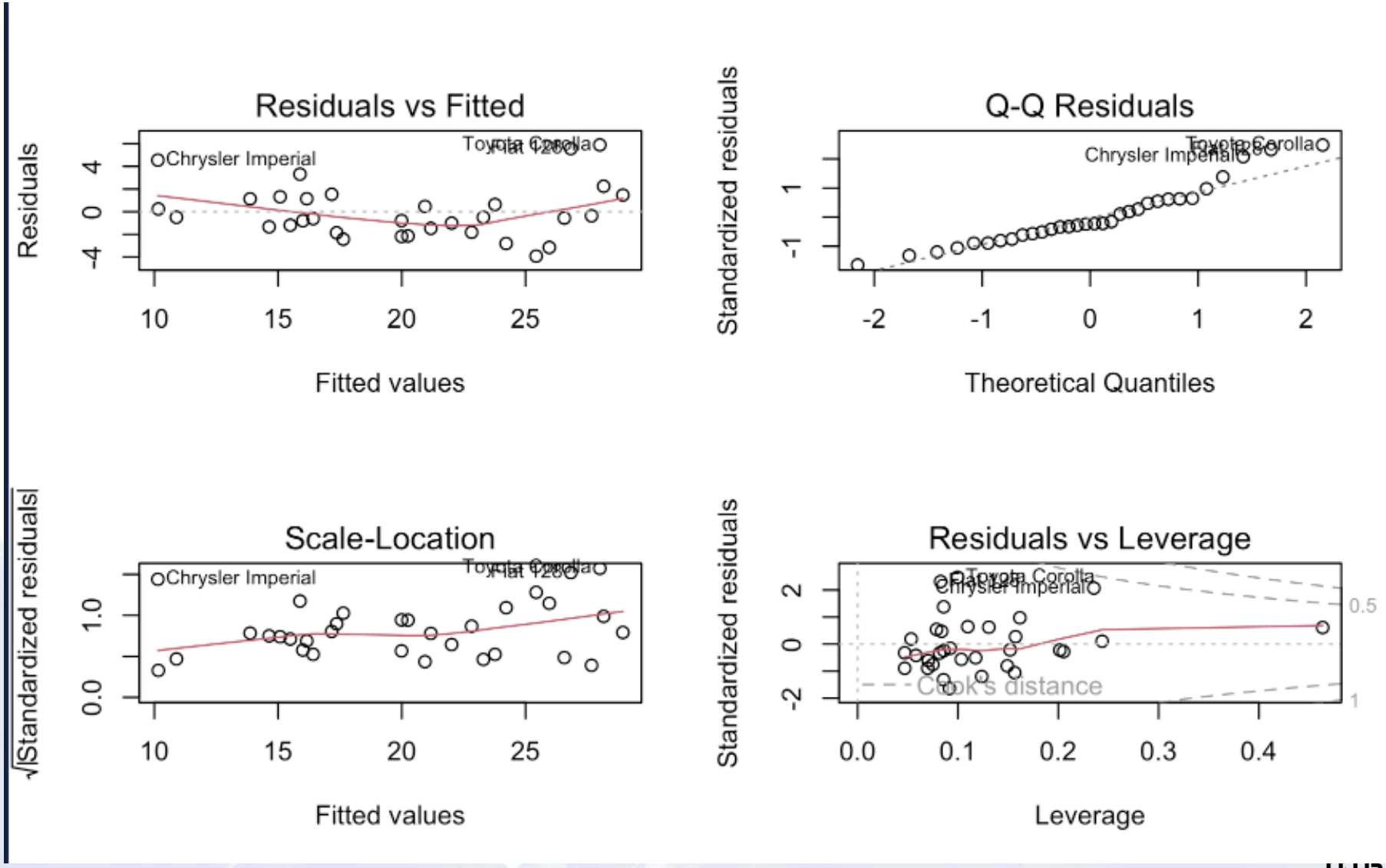
Aufgaben lösen

Interpretation der Ergebnisse:

- Erklären Sie die geschätzten Koeffizienten und die Bedeutung der einzelnen Prädiktoren.
- Identifizieren Sie, welche Variablen signifikant sind und welche nicht.
- Erklären Sie, was der F-Wert und R^2 über das Modell aussagen.

Model-summary

provadis Hochschule



Model-summary

Res vs. fitted:

- **Zweck:**

Dieser Plot zeigt die Residuen (Fehler) des Modells gegen die vorhergesagten Werte (fitted values).
Er hilft, die Annahme der Homoskedastizität (konstante Varianz der Fehler) und die Linearität zu überprüfen.

- **Interpretation:**

Die Residuen sollten zufällig um die Nulllinie streuen, ohne erkennbare Muster.
Ein klarer Trend oder ein Muster (wie eine Parabel) deutet darauf hin, dass das Modell möglicherweise nicht alle Trends in den Daten erfasst oder dass eine Nichtlinearität vorliegt

QQ plot:

- **Zweck:**

Der QQ-Plot vergleicht die Verteilung der Modellresiduen mit einer theoretischen Normalverteilung.
Er überprüft die Annahme der Normalverteilung der Fehler.

- **Interpretation:**

Wenn die Residuen normalverteilt sind, liegen die Punkte nahe an der Diagonale (45-Grad-Linie).
Große Abweichungen von der Linie deuten auf Abweichungen von der Normalverteilung hin.

#Scale location:

- **Zweck:**

Dieser Plot zeigt die quadrierte Wurzel der standardisierten Residuen gegen die vorhergesagten Werte.
Er hilft, die Homoskedastizität der Residuen zu überprüfen.

- **Interpretation:**

Die Punkte sollten gleichmäßig verteilt sein, ohne systematische Muster.
Ein Trend oder eine Trichterform (Zunahme oder Abnahme der Streuung) deutet auf Heteroskedastizität hin.

Cook's distance:

- **Zweck:**

Cook's Distance misst den Einfluss jedes Datenpunkts auf die Gesamtschätzung des Modells.
Es identifiziert potenziell einflussreiche Datenpunkte (Outlier).

- **Interpretation:**

Punkte mit einem hohen Cook's Distance-Wert (> 0.5 oder 1) können als einflussreich angesehen werden und sollten untersucht werden.
Solche Punkte könnten das Modell unverhältnismäßig beeinflussen und müssen möglicherweise überprüft oder entfernt werden.

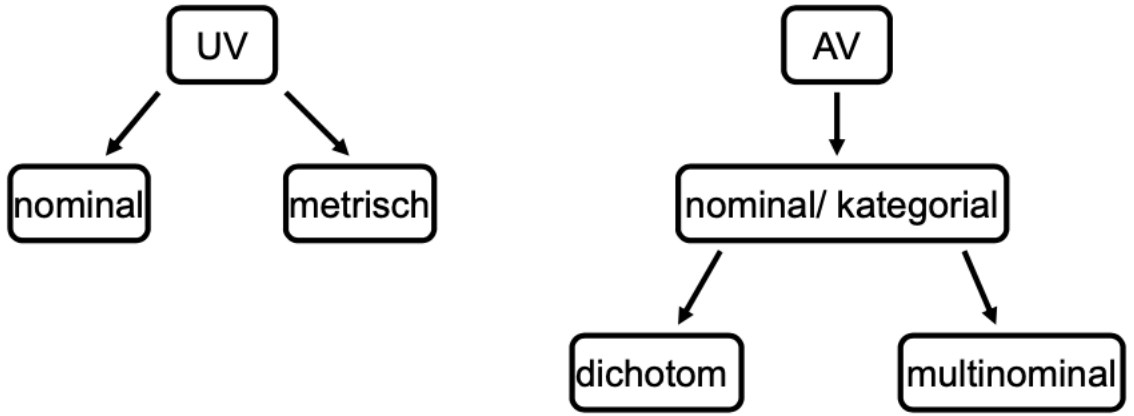
Logistische Regression: Definition

- Die logistische Regression ist ein statistisches Verfahren zur Vorhersage des Wertes einer dichotomen abhängigen Variable (z.B. ja/nein, wahr/falsch, 0/1) basierend auf einer oder mehreren unabhängigen Variablen.
- Dichotome Variablen sind Variablen, die nur zwei mögliche Werte haben.
- Im Gegensatz zur linearen Regression, die kontinuierliche Ergebnisse liefert, gibt die logistische Regression Wahrscheinlichkeiten für die Kategorien der binären abhängigen Variable aus.
- Probleme mit mehr als zwei Kategorien werden als multinomiale logistische Regression oder, wenn die mehreren Kategorien geordnet sind, als geordnete logistische Regression bezeichnet.

Logistische Regression: Definition

Warum die Logistische Regression?

- **Medizin:** Vorhersage, ob ein Patient eine bestimmte Krankheit hat (ja/nein) basierend auf Symptomen und Testergebnissen.
- **Marketing:** Vorhersage, ob ein Kunde ein Produkt kaufen wird (ja/nein) basierend auf demografischen Daten und Kaufverhalten.
- **Sozialwissenschaften:** Vorhersage, ob ein Ereignis stattfindet (z.B. Wahlteilnahme) basierend auf Umfragedaten.
- **Politik**: wählen einer Partei A / nicht
- **Prüfung**: Bestanden / nicht bestanden



Logistische Regression: Definition

provadis Hochschule

Warum brauchen wir logistische Regression?

- Klassische **lineare Regression** eignet sich für kontinuierliche Zielvariablen.
- **Klassifikationsprobleme** (z. B. Spam vs. Kein Spam, Krank vs. Gesund) erfordern eine andere Herangehensweise.
- Ziel: **Wahrscheinlichkeit für eine Klasse** vorhersagen.
- Problem: Lineare Modelle liefern Werte außerhalb des Bereichs [0,1].

Logistische Vs. Lineare Regression

- Vorhersage einer AV mittels einer oder mehreren UV
- !! AV nicht metrisch sondern nominal
- !! Keine Vorhersage der tatsächlichen Werten sondern deren Eintrittswahrscheinlichkeit
- S-förmiger Verlauf
- Wahrscheinlichkeit für Y = 1 liegt im Intervall [0,1]
- symmetrisch um Wendepunkt P(y = 1) = 0,5

Logistische Vs. Lineare Regression

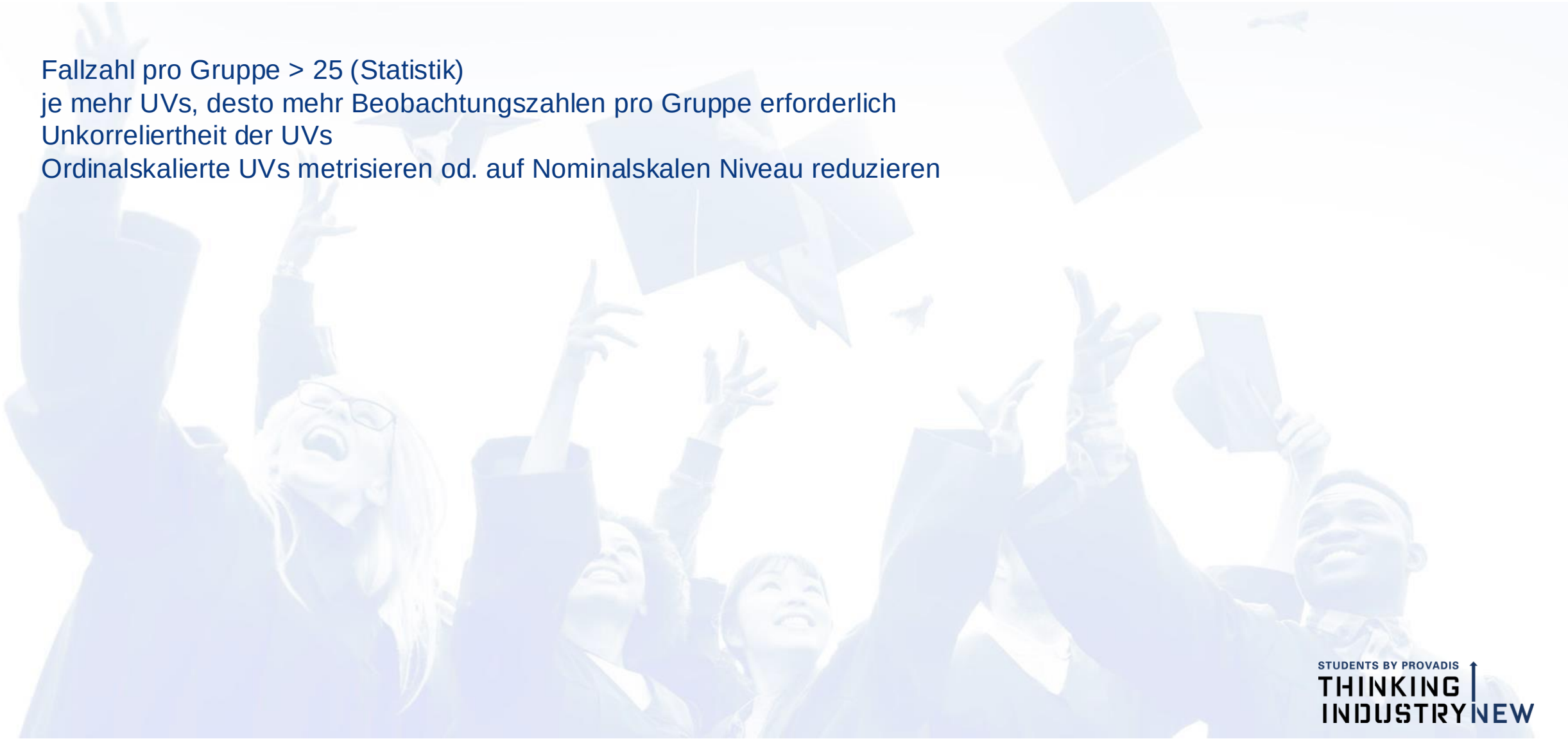
provadis Hochschule

- Lineare Modelle:
  - Modell: $y = ß_0 + ß_1 x$
  - Ausgabe: Beliebige Werte $(-\infty, +\infty)$
  - Nicht geeignet für Klassifikationen

- Logistische Modelle:
  - Modelliert **Wahrscheinlichkeiten** im Bereich [0,1]
  - Verwendet **Sigmoid-Funktion** zur Transformation

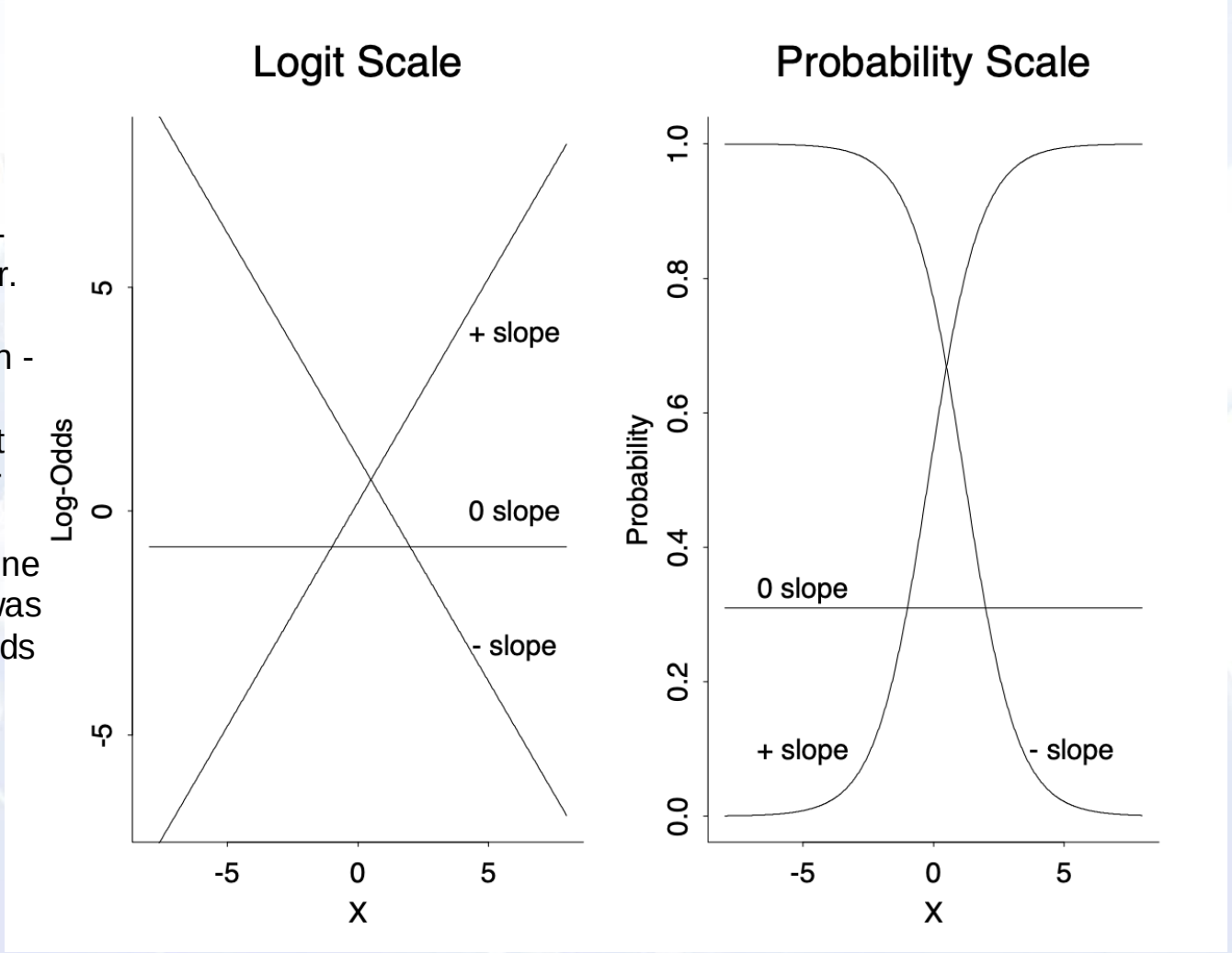
Logistische Regression: Anforderung

Fallzahl pro Gruppe > 25 (Statistik)
je mehr UVs, desto mehr Beobachtungszahlen pro Gruppe erforderlich
Unkorreliertheit der UVs
Ordinalskalierte UVs metrisieren od. auf Nominalskalen Niveau reduzieren



Einfaches Modell

- Die Logit-Skala stellt die lineare Kombination (Log-Odds) der Prädiktoren dar.
- Die X-Achse zeigt den Bereich der Log-Odds von -5 bis +5.
- Die Y-Achse repräsentiert ebenfalls den Bereich der Log-Odds von -5 bis +5.
- Jede Linie im Plot zeigt eine verschiedene Steigung, was die Änderung der Log-Odds darstellt.



- Die Wahrscheinlichkeitsskala zeigt S-förmige Kurven, die die Wahrscheinlichkeiten (P) der Ereignisse darstellen.
- Die X-Achse repräsentiert den Bereich der Log-Odds von -5 bis +5.
- Die Y-Achse zeigt die Wahrscheinlichkeit (P) von 0 bis 1.
- Jede Kurve hat eine S-förmige Form mit verschiedenen Steigungen.

Log-Odds (Logit Funktion)

Anstatt die Wahrscheinlichkeit ppp direkt vorherzusagen, nutzt die logistische Regression das Logit-Transform:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = ß_0 + ß_1 x_1 + \cdots + ß_n x_n$$

Der Ausdruck **Log-Odds** bedeutet "logarithmierte Chancen".

Wenn $\log\left(\frac{p}{1-p}\right) = 0$, dann ist p=0.5, weil die Chancen 1:1 sind.

Wenn der Wert stark positiv ist, geht p gegen 1, wenn er stark negativ ist, geht p gegen 0.

Die Sigmoid-Funktion

provadis Hochschule

- Die Logistische Funktion (Sigmoid) ist definiert als:

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^{z}}{1 + e^{z}}$$

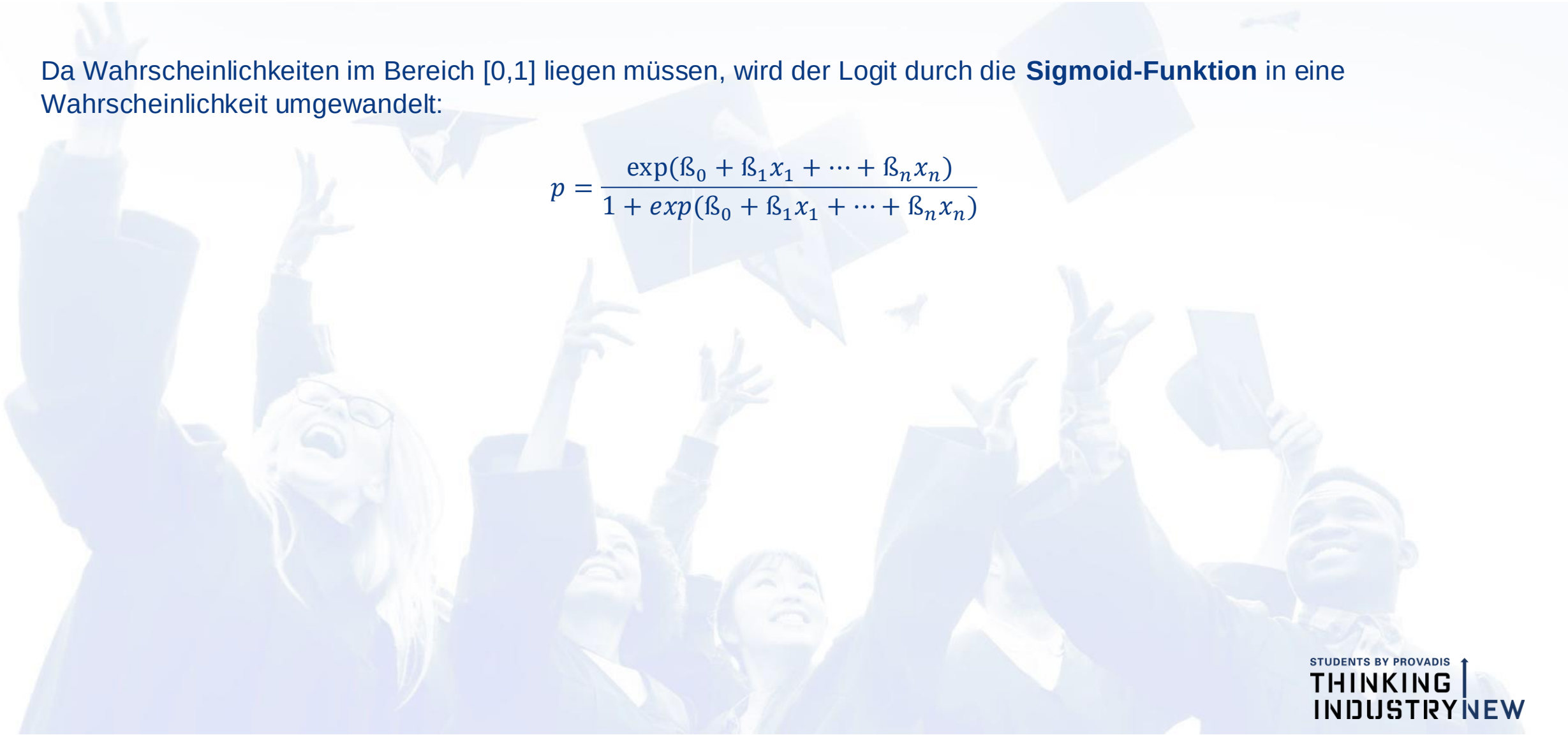
Wobei:

Z= $ß_0 + ß_1 x$
•Wertebereich: [0,1]
•Für große Z →1 , für kleine Z→0
•Glatte, nicht-lineare Transformation

Die Sigmoid-Funktion

Da Wahrscheinlichkeiten im Bereich [0,1] liegen müssen, wird der Logit durch die **Sigmoid-Funktion** in eine Wahrscheinlichkeit umgewandelt:

$$p = \frac{\exp(ß_0 + ß_1 x_1 + \cdots + ß_n x_n)}{1 + exp(ß_0 + ß_1 x_1 + \cdots + ß_n x_n)}$$



Multinomiale Logistische Regression –Softmax Reg

Wird verwendet, wenn es **mehr als zwei** Klassen gibt, die **keine natürliche Ordnung** haben (z. B. Klassifizierung von Tieren: Katze, Hund, Vogel).

Verallgemeinert die **binäre logistische Regression** mit der **Softmax-Funktion**.

Softmax-Funktion
Anstatt nur eine Wahrscheinlichkeit für eine Klasse zu berechnen, berechnen wir für jede Klasse kkk eine Wahrscheinlichkeit:

$$p_k = \frac{e^{ß_{0k}+\sum_{j=1}^{p} ß_{jk}x_j}}{\sum_{l=1}^{k} e^{ß_{0k}+\sum_{j=1}^{p} ß_{jk}x_j}}$$

Multinomiale Logistische Regression –Softmax Reg

provadis Hochschule

Softmax-Funktion

$$p_k = \frac{e^{ß_{0k} + \sum_{j=1}^{p} ß_{jk} x_j}}{\sum_{l=1}^{K} e^{ß_{0l} + \sum_{j=1}^{p} ß_{jl} x_j}}$$

k: Die aktuelle Klasse, für die wir die Wahrscheinlichkeit berechnen.
p_k: Wahrscheinlichkeit für Klasse k / wobei Klasse l:1 bis k.
β_{0k}: Intercept (Bias) für Klasse k.
β_{jk}: Koeffizienten für Variable x_j für Klasse k.
p: Anzahl der erklärenden Variablen (z. B. Gewicht, Größe, Alter).
K: Gesamtanzahl der Klassen.

Multinomiale Logistische Regression –Softmax Reg

provadis Hochschule

Softmax-Funktion

$$p_k = \frac{e^{ß_{0k}+\sum_{j=1}^{p} ß_{jk}x_j}}{\sum_{l=1}^{k} e^{ß_{0k}+\sum_{j=1}^{p} ß_{jk}x_j}}$$

Numerator: Exponentieller Wert des linearen Modells für Klasse k.
Denominator: Summe aller exponentiellen Werte für **alle** Klassen.
Dadurch werden die Wahrscheinlichkeiten so skaliert, dass ihre Summe **1** ergibt.

Wie wird die Wahrscheinlichkeit berechnet?

- Wir berechnen für jede Klasse k den **exponentiellen Wert der linearen Funktion**.
- Danach teilen wir durch die **Summe aller exponentiellen Werte für alle Klassen**, um sicherzustellen, dass die Wahrscheinlichkeiten insgesamt **1 ergeben**.

→ Die Klasse mit der höchsten Wahrscheinlichkeit wird vorhergesagt.
→ Ähnlich zur binären logistischen Regression, aber statt der **Sigmoid-Funktion** wird **Softmax** verwendet.

provadis Hochschule

Beispiel (Tiere klassifizieren: Katze, Hund, Vogel)

Angenommen, wir haben 3 Klassen:

- **Katze (k = 1)**
- **Hund (k = 2)**
- **Vogel (k = 3)**

Nehmen wir an, die berechneten Werte vor der Softmax-Transformation sind:

- Katze: $e^{2.5}=12.18$
- Hund: $e^{1.2}=3.32$
- Vogel: $e^{0.8}==2.23$

Multinomiale Logistische Regression –Softmax Reg

provadis Hochschule

Beispiel (Tiere klassifizieren: Katze, Hund, Vogel)

$$p_{Katze} = \frac{12.18}{12.18 + 3.32 + 2.23} = 0.67$$

$$p_{Hund} = \frac{3.32}{12.18 + 3.32 + 2.23} = 0.18$$

$$p_{Vogel} = \frac{2.23}{12.18 + 3.32 + 2.23} = 0.15$$

Die Klasse mit der **höchsten Wahrscheinlichkeit (Katze, 67%)** wird vorhergesagt.

Ordinale Logistische Regression

Wird verwendet, wenn die Klassen eine **natürliche Ordnung** haben (z. B. Bewertungen: "schlecht", "mittel", "gut").

Statt für jede Klasse eine eigene Funktion zu modellieren, wird eine **kumulative Wahrscheinlichkeit** genutzt.

Kumulative Logit-Funktion (Proportional Odds Model)

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \alpha_k - \sum_{j=1}^{p} ß_j x_j$$

•Modelliert **kumulative Wahrscheinlichkeiten**, also die Wahrscheinlichkeit, dass eine Beobachtung in Klasse **≤k** liegt.

Ordinale Logistische Regression

Kumulative Logit-Funktion (Proportional Odds Model)

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \alpha_k - \sum_{j=1}^{p} ß_j x_j$$

- P(Y≤k) die Wahrscheinlichkeit ist, dass die Klasse **höchstens** k ist.
- $\alpha_k$sind die **Schwellenwerte** (Intercepts) für jede Kategorie k.
- β_j sind die **Regressionskoeffizienten** für die Prädiktoren X_j.
- p ist die Anzahl der erklärenden Variablen.

Ordinale Logistische Regression

Wahrscheinlichkeitsberechnung:

$$P(Y \leq k) = \frac{1}{1 + e^{-(\alpha_k - \sum_{j=1}^{p} ß_j x_j)}}$$

Die Wahrscheinlichkeit für eine bestimmte Klasse P(Y=k) ergibt sich dann durch Differenzen:

$$P(Y = k) = P(Y \leq k) - P(Y \leq k - 1)$$

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

Wir betrachten eine Umfrage zur **Kundenzufriedenheit** mit einem Produkt. Die Kunden bewerten das Produkt auf einer Skala von 1 bis 5:

1 = „Sehr schlecht“, 2 = „Schlecht“, 3 = „Mittel“, 4 = „Gut“, 5 = „Sehr gut“.

Die geschätzten Schwellenwerte α_k und Koeffizienten β_j könnten z. B. sein:

- $\alpha 1 = -2.0$
- $\alpha 2 = -0.5$
- $\alpha 3 = 1.0$
- $\alpha 4 = 2.5$
- $\beta 1 = 1.2$ (z. B. für einen Zufriedenheitsfaktor wie Produktqualität)

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

Das bedeutet:
•Ein **höherer** Produktqualitätswert X erhöht die Wahrscheinlichkeit für eine **höhere** Zufriedenheitsklasse.

•Die Schwellenwerte bestimmen, wo die Wahrscheinlichkeiten zwischen den Klassen wechseln.

Ein Kunde gibt eine Produktqualität von X=4 :

1- Wahrscheinlichkeit für Y≤1 (Sehr schlecht oder schlechter):

$$P(Y \leq k) = \frac{1}{1 + e^{-(\alpha_k - ßX)}}$$

→

$$P(Y \leq 1) = \frac{1}{1+e^{-(-2-(1.2*4))}} \approx 0.0011$$

→ **Fast 0% Wahrscheinlichkeit für „Sehr schlecht“.**

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

2- Wahrscheinlichkeit für Y≤2 (schlecht oder schlechter):

→

$$P(Y \leq 2) = \frac{1}{1+e^{-(-0.5-(1.2*4))}} \approx 0.013$$

→ **1.3% Wahrscheinlichkeit für „schlecht“ oder „schlechter“.**

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

3- Wahrscheinlichkeit für Y≤3 (Mittel oder schlechter):

→

$$P(Y \leq 3) = \frac{1}{1+e^{-(1-(1.2*4))}} \approx 0.0909$$

→ **9.1% Wahrscheinlichkeit für „Mittel“ oder „schlechter“.**

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

4- Wahrscheinlichkeit für Y≤4 (Gut oder Schlechter):

→

$$P(Y \leq 4) = \frac{1}{1 + e^{-(2.5-(1.2*4))}} \approx 0.332$$

→ **33.2% Wahrscheinlichkeit für „ Gut“ oder „ schlechter“.**

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

$$P(Y = k) = P(Y \leq k) - P(Y \leq k - 1)$$

Wahrscheinlichkeiten für jede Klasse:

- Sehr schlecht: P(Y=1)=P(Y<=1)= 0.0011
- Schlecht: P(Y=2)=P(Y<=2) - P(Y<=2-1) = 0.0013 - 0.0011 = 0.0119
- Mittel: P(Y=3)= P(Y<=3) - P(Y<=2)= 0.0909 – 0.012 = 0.078
- Gut: P(Y=4)= P(Y<=4) - P(Y<=3)= 0.332 – 0.0909 = 0.241
- Sehr Gut: P(Y=5)= P(Y<=5) - P(Y<=4)= 1 – 0.332 =0.668

provadis Hochschule

Beispiel: Einstufung der Kundenzufriedenheit

Fazit für diesen Kunden mit Produktqualität X=4
Die Wahrscheinlichkeiten für die Zufriedenheitsstufen sind:

- **„Sehr schlecht“**: **0.11%**
- **„Schlecht“**: **1.19%**
- **„Mittel“**: **7.8%**
- **„Gut“**: **24.1%**
- **„Sehr gut“**: **66.8%**

Interpretation:

•Der Kunde mit einer Produktqualität von **4 von 10** hat eine **hohe Wahrscheinlichkeit (66.8%) für „Sehr gut“**, aber auch eine gewisse Wahrscheinlichkeit für „Gut“ (24.1%).
•Die Wahrscheinlichkeit für „Schlecht“ oder „Sehr schlecht“ ist **sehr gering (unter 2%)**.
•Das zeigt, dass **Produktqualität stark mit Zufriedenheit korreliert**, aber auch, dass selbst mittlere Qualität noch eine Chance auf hohe Bewertungen hat.

Ordinale Logistische Regression

provadis Hochschule

Methode	Anwendung	Wahrscheinlichkeit durch
Binäre logistische Regression	2 Klassen	Sigmoid-Funktion
Multinomiale Regression	>2 ungeordnete Klassen	Softmax-Funktion
Ordinale Regression	>2 geordnete Klassen	Kumulative Logit-Funktion

Zurück zur binären Logreg- Sigmoid Fct.
Maximum-Likelihood-Schätzung

Statt der MSE (wie bei linearer Regression) wird die **Likelihood maximiert**:

$$L(ß) = \prod_{i=1}^{n} P(Y_i|Xi)$$

oder logarithmiert:

$$logL(ß) = \sum_{i=1}^{n} [y_i \log P(Y_i) + (1 - y_i)\log(1 - P(Y_i))]$$

Ziel: Finden von , die die Daten am besten erklären.

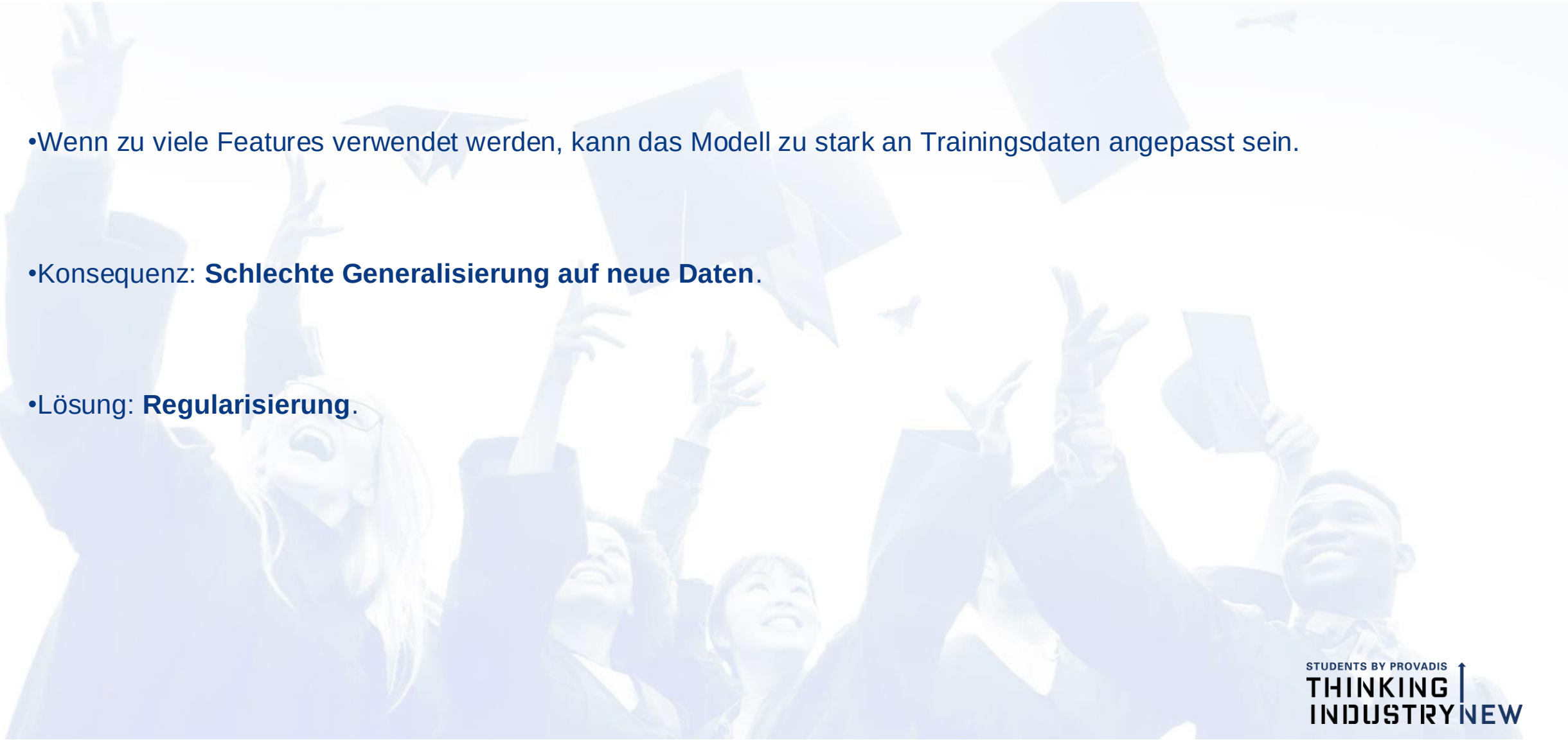
Problem bei der Überanpassung

provadis Hochschule

•Wenn zu viele Features verwendet werden, kann das Modell zu stark an Trainingsdaten angepasst sein.

•Konsequenz: **Schlechte Generalisierung auf neue Daten**.

•Lösung: **Regularisierung**.



Lasso Regression (L1-Regularisierung)

provadis Hochschule

- Fügt eine **L1-Strafe** hinzu:

$$J(ß) = -\log L(ß) + \lambda \sum_{j=1}^{p} |ß_j|$$

- Kann einige Koeffizienten **genau auf 0 setzen** → Feature Selection
- Hilft, irrelevante Variablen zu eliminieren
- Gut für **sparsame Modelle**
- Die Lösung kann **nicht analytisch** berechnet werden, sondern wird z. B. mit **Coordinate Descent** oder **Gradient Descent** bestimmt.

Ridge Regression (L2-Regularisierung)

provadis Hochschule

• Fügt eine **L2-Strafe** hinzu:
Zielfunktion:

$$J(ß) = -\log L(ß) + \lambda \sum_{j=1}^{p} ß_j^2$$

- Verkleinert Koeffizienten, aber setzt sie **nicht auf 0**
- Stabilisiert das Modell bei **hoher Multikollinearität**
- Gut für Modelle, bei denen alle Features relevant sind

Vergleich Lasso Vs. Ridge

provadis Hochschule

Regularisierungstyp:

Lasso-Regression verwendet die absoluten Koeffizienten in der Regularisierung und hat die Eigenschaft, einige Koeffizienten auf Null zu setzen, was eine Art von Variablenauswahl ermöglicht.

Ridge-Regression verwendet die quadratischen Koeffizienten in der Regularisierung, was dazu führt, dass große Koeffizienten reduziert werden.

Vergleich Lasso Vs. Ridge

provadis Hochschule

Koeffizienten:

Ridge-Regression werden die Koeffizienten zwar reduziert, aber normalerweise nicht auf Null gesetzt.

In Lasso können einige Koeffizienten auf Null gesetzt werden, was zu einem spärlichen Modell führt.

Vergleich Lasso Vs. Ridge

provadis Hochschule

Komplexität:

Lasso-Regression führt zu spärlicheren Lösungen und kann daher zur Variablenauswahl verwendet werden, wodurch das Modell einfacher und interpretierbarer wird.

Da J(ß) nicht analytisch lösbar ist (außer für Ridge in linearen Modellen), nutzen wir **Gradient Descent**.

Ridge-Regression führt zu glatteren Lösungen und behält normalerweise alle Variablen im Modell bei, reduziert jedoch die Einflussstärke.

Lasso Schätzer :

$$ß_j^* = sign\, ß_j . \max(0, |ß_j| - \lambda)$$

Ridge Schätzer:

$$ß_j^* = \frac{ß_j}{1+\lambda}\ Alle\ Koeff - gleichmäßig\ verkleinern$$

Vergleich Lasso Vs. Ridge

Eigenschaft	Lasso (L1)	Ridge (L2)
Bestrafung	$\sum \lvert ß_j \rvert$	$\sum ß_j^2$
Feature Selection	Ja (einige)	Nein (alle)
Effekt auf Koeffizienten	Setzt manche auf genau 0	Reduziert Koeffizienten, aber keine 0
Einsatzbereich	Wenige wichtige Features	Viele kleine Effekte
Analytisch lösbar?	Nein	Ja für Lineare Modelle

Elastic Net: L-1 & L-2 Kombination

Ziel Funktion:

$$J(ß) = -\log L(ß) + \lambda_1 \sum |ß_j| + \lambda_2 \sum ß_j^2$$

Eigenschaften:

- Gut für **viele korrelierte Features**
- Nutzt sowohl Feature Selection als auch Stabilisierung

ElasticNet Schätzer:

$$ß_j^* = \frac{sign(ß_j).\max(0, |ß_j| - \lambda_1)}{1 + \lambda_2}$$

Lasso - Ridge - Elastic Net

provadis Hochschule

- Lasso: Manche Koeffizienten verschwinden
- Ridge: Alle Koeffizienten werden kleiner
- Elastic Net: Mischung aus beiden Effekten

- **Kreuzvalidierung** wird genutzt, um zu optimieren.

- Je größer :
  - Stärkere Regularisierung
  - Mehr Bias, weniger Varianz
- Je kleiner :
  - Weniger Regularisierung
  - Weniger Bias, mehr Varianz

Bsp-

Angenommen, wir haben ein Modell mit 3 Features und die **OLS-Koeffizienten** wären:

- $ß_1 = 5$
- $ß_2 = -3$
- $ß_3 = 0.5$

Wir fügen nun Regularisierungen hinzu.
Die Berechnung erfolgt durch Minimierung der jeweiligen Kostenfunktion.

provadis Hochschule

LASSO

Lasso fügt eine **L1-Strafe** hinzu:

$$J(ß) = -\log L(ß) + \lambda \sum_{j=1}^{p} |ß_j|$$

Dadurch können einige Koeffizienten auf **genau 0** gesetzt werden.
Für ein Beispiel mit $\lambda = 1$:

Berechnung der Lasso- Schätzern:

$$ß_j^* = sign\, ß_j . \max(0, |ß_j| - \lambda)$$

LASSO

Bsp. $\lambda = 1$:

$ß_1^* = sign(5).\max(0, |5| - 1)$=4
$ß_2^* = sign(-3).\max(0, |-3| - 1)$= -2
$ß_3^* = sign(0.5).\max(0, |0.5| - 1)$=0 → wird entfernt

Bsp. $\lambda = 10$:

$ß_1^* = sign(5).\max(0, |5| - 10)$=0
$ß_2^* = sign(-3).\max(0, |-3| - 10)$= 0
$ß_3^* = sign(0.5).\max(0, |0.5| - 10)$=0

Bsp-

RIDGE

RIDGE fügt eine **L2-Strafe** hinzu:

$$J(ß) = -\log L(ß) + \lambda \sum_{j=1}^{p} |ß_j|$$

Dadurch werden **alle Koeffizienten verringert**, aber nicht auf 0 gesetzt.

Berechnung der Ridge Schätzern → Ableitung der Zielfunktion

$$ß_j^* = \frac{ß_j}{1+\lambda}$$

Bsp-

RIDGE

Bsp. $\lambda = 1$:

$$ß_1^* = \frac{5}{1+1} = 2.5$$

$$ß_2^* = \frac{-3}{1+1} = -1.5$$

$$ß_2^* = \frac{0.5}{1+1} = 0.25$$

Bsp. $\lambda = 10$:

$$ß_1^* = \frac{5}{1+10} = 0.45$$

$$ß_2^* = \frac{-3}{1+10} = -0.27$$

$$ß_2^* = \frac{0.5}{1+10} = 0.045$$

provadis Hochschule

Elastic-Net

Elastic Net kombiniert die vorherigen Methoden:

$$J(ß) = -\log L(ß) + \lambda_1 \sum |ß_j| + \lambda_2 \sum ß_j^2$$

Berechnung der ElasticNet Schätzern → Ableitung der Zielfunktion

$$ß_j^* = \frac{sign(ß_j).\max(0, |ß_j| - \lambda_1)}{1 + \lambda_2}$$

Elastic-Net

Bsp. $\lambda_1 = 0.5\ und\ \lambda 2 = 0.5$:

$$ß_1^* = \frac{sign(5).\max(0, |5| - 0.5)}{1 + 0.5} = 3$$

$$ß_2^* = \frac{sign(-3).\max(0, |-3| - 0.5)}{1 + 0.5} = -1.67$$

$$ß_3^* = \frac{sign(0.5).\max(0, |0.5| - 0.5)}{1 + 0.5} = 0$$

Zusammenfassung

**Wenn ** → Keine Regularisierung, Standard-OLS.

- **Wenn sehr groß ist**:
  - **Lasso** entfernt fast alle Koeffizienten (Feature Selection).
  - **Ridge** reduziert Koeffizienten stark, aber nie auf 0.
  - **Elastic Net** kombiniert beide Effekte und kann je nach Parametern zwischen Ridge- und Lasso-Verhalten wechseln.

Durch die Wahl von kann man den Grad der Regularisierung steuern und das Modell an die Daten anpassen.

Vielen Dank für ihre Aufmerksamkeit!

Bei Fragen nehmen Sie gerne Kontakt auf:

Provadis School of International Management and Technologies AG

Dr. Houssam Jedidi

Industriepark Höchst Frankfurt am Main

Tel.: +49 176 30453606

Mail: houssam.jedidi@doz-provadis-hochschule.de