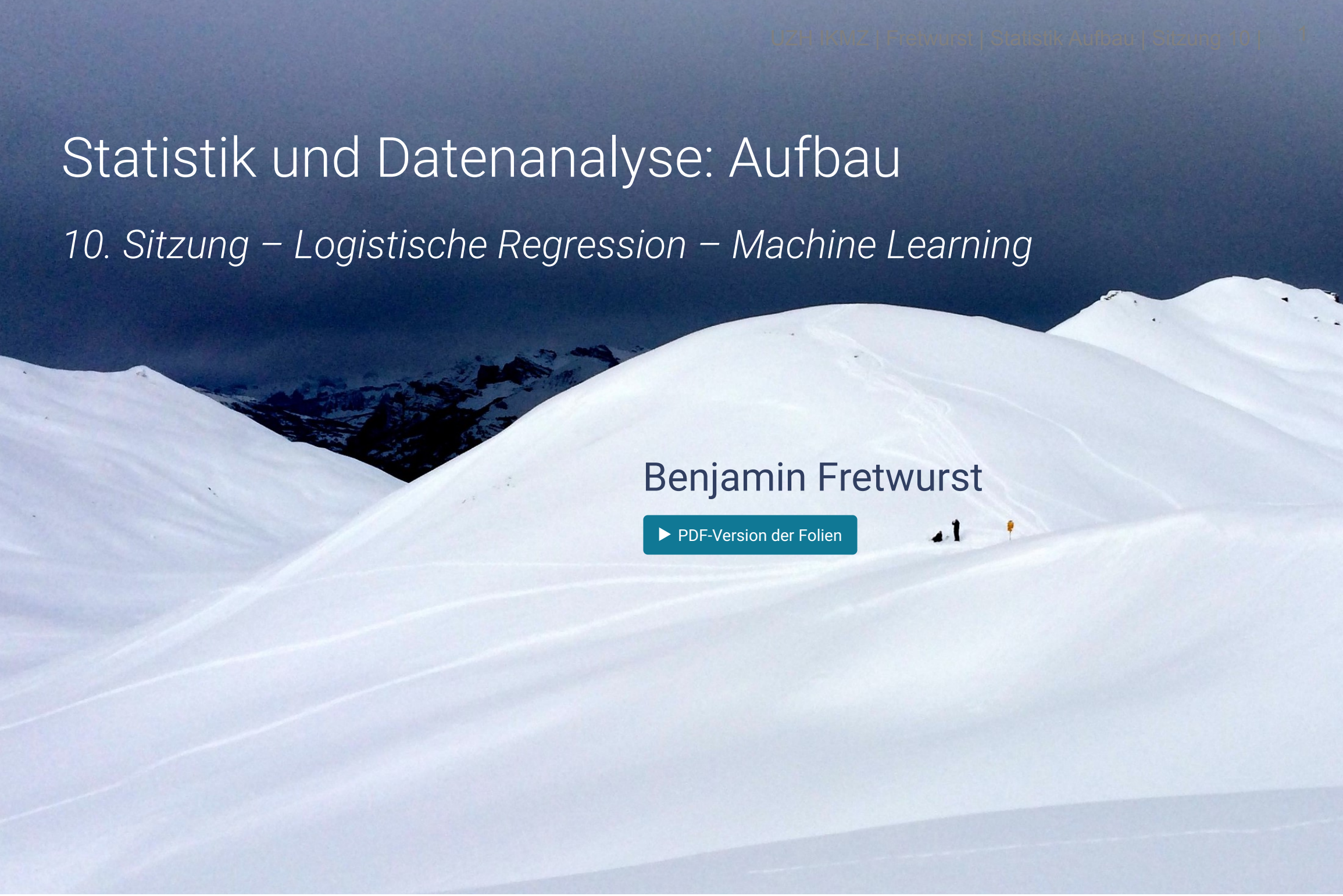
Statistik und Datenanalyse: Aufbau

10. Sitzung – Logistische Regression – Machine Learning



Benjamin Fretwurst

▶ PDF-Version der Folien

Inhalt

Orga

Foto: Lisa Vogel

Klausur

Anlage

- Die Klausur basiert zu mind. 60% auf den LEF und zu 100% auf deren Flughöhe
- Es wird ca. 20 MC-Fragen a 2P gegben und 20P für 4–5 Essayfragen.
- Es kann auch R-Output und/oder Tabellen aus Studien.
- LEF enthalten keine Fehler, die Ihnen angekreidet würden

Punktverteilung (± 2) MC nach Bereichen von 60:

- Regression (linear, kurvlienear, Interaktion, Kennwerte, Interpretation) 10P
- Voraussetzungen OLS 8P
- Dummys und Slope-Dummys UVs 6P
- R (pauschal) 2P
- Faktorenanalyse 8P
- Logistische Regression 4P
- Clusteranalyse 4P

Essayfragen Fragen zu Grundlagen aller Bereiche und Tabellenoutputs 20P

Lernziele

Theoretisch

- Festigung GLM
- Grundprinzip Machine Learning (ML)
- Logistische Regression
- Multinominale Regression
- Zum nächsten Mal Übung 4 und Text Zerback Wirz! lesen! (liegt beides auch in OLAT unter Materialien bzw. Texte)

1 Regression multivariat

1.1 Das Modell mit 2 UV's

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U \quad (1)$$

$$Y_i = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \quad (2)$$

Elemente der Gleichng

- (1) ist das theoretische Modell und (2) die Schätzgleichung
- die βs sind unbekannte Parameter (darum griechisch)
- das U ist der unbekannte Rest an Einflüssen
- die Y und X sind Variablen der Erhebung, aber bei der Analyse fixe Daten
- die b's müssen anhand der Stichprobendaten berechnet werden und schätzen die βs
- die b's sind skalenabhängig und die $BETA = b\frac{s_x}{s_y}$ nicht
- b's und BETAs werden mit t-Tests auf Signifikanz geprüft (p-value < .05)
- Enthalten die CI's der b's (oder BETA's) *nicht* die 0, sind die Kennwerte signifikant

1.2 Voraussetzungen und Probleme

die b's sind BLUE (Best Linear Unbiased Estimator), wenn:

1. X und Y sind fix
2. das Modell voll spezifiziert ist, also keine Einflussgrössen fehlen ($r_{Ue_i} = 0$)
3. die Residuen homoskedastisch sind, also überall gleich um die $\hat{Y}$ streuen
4. keine Autokorrelation der Residuen (gibts nur bei longitudinalen Daten)

Weitere Probleme

- bei perfekter Multikollinearität kann nicht gerechnet werden
- bei hoher Multikollinearität (Tol < .5; VIF > 2), sind die b's sehr unsicher (se_b gross)
- sind die Residuen nicht (multivariat) normalverteilt, ist der t-Test nicht zuverlässig
- Ausreisser (Outlier) beeinflussen die b's stark.

1.3 Identifikation und Lösungen

Modellspezfikation

Es braucht eine gute Theorie und Arbeit an Modellspezifikation

Heteroskedastizität

Identifikation: Residualplots und Homoskedastizitätstests

Lösung:

- Transformation bei nichtlinearen Beziehungen
- 2-SLS

Residuen nicht normalverteilt

Identifikation: Residualplots und NV-Test
Lösung: Bootstrapping

Multikollinearität

Identifikation: TOL und VIF
Lösung: Faktorenanalyse → Variablenausschluss oder Indizes

Outlier

Identifikation: Residuals vs Leverage plot
Lösung: Outlier raus / robuste Regression

2 Machine Learning – ML

Unsupervised Learning

Beim Unsupervised Learning versucht die «Maschiene» ohne Vorgaben, eigenständig Strukturen in den Daten zu finden. Faktorenanalysen (PCA) wird beim Preprocessing der Daten eingesetzt und Clusteranalysen, um Gruppen zu identifizieren.

Supervised Learning

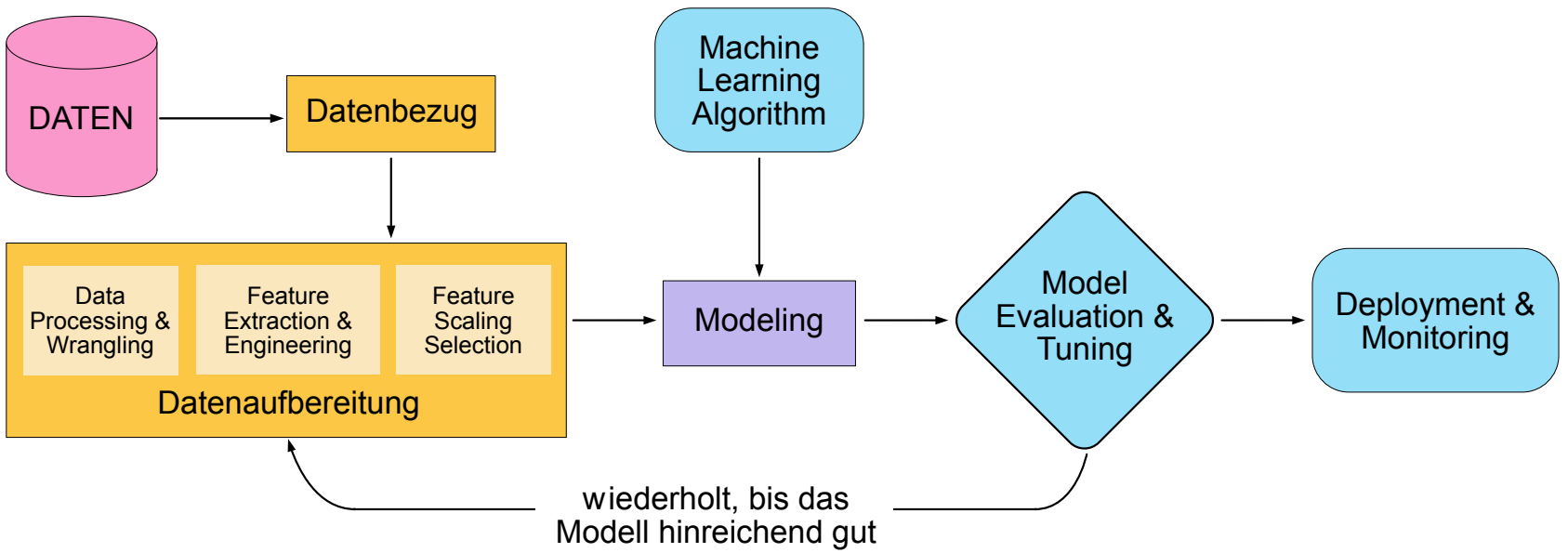
Beim Supervised Learning werden der Maschiene Vorgaben gemacht, indem ihr Trainingsdaten gegeben werden, die Erkennungsmerkmale («Feature» aka UVs) und Ergebnis enthalten. Die Maschiene sucht nach Modellen, die versucht aus den «Features» das Ergebnis abzuleiten. Die Challenge besteht darin, gute Feature zu finden und wird «Feature Engineering» genannt.

Supervised

Trainieren ➪ Testen ➪ Anwenden

Die verfügbaren Daten werden in einen grösseren und einen kleineren Teil getrennt:

1. Trainingsdaten, an denen das Modell trainiert wird (Modelling)
2. Testdaten, an denen das Modell evaluiert wird.



3 Logistische Regression

3.1 Grund und Folgen

Warum?!

Wir haben nicht immer schöne metrische Variablen. Manchmal wollen wir dichotome Merkmale erklären oder kategoriale. Dann werden logistische Regressionen angewendet.

Wir schätzen mit einem logistischen Modell die Eintrittswahrscheinlichkeit einer dichotomen AV.

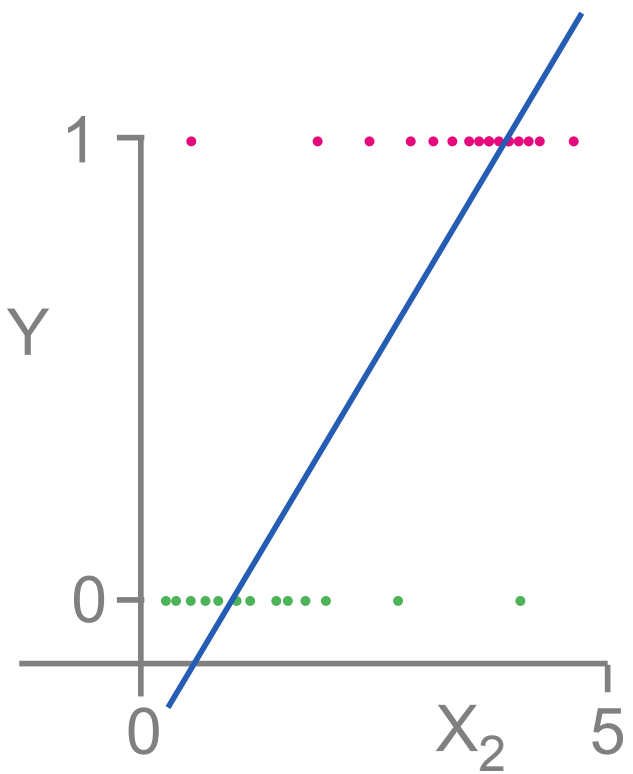
Logistische Regression hat das gleiche Ziel wie Diskriminanzanalysen. Wir können Abhängige diskriminieren. Klingt schlimm, ist aber technisch gemeint.

Beispiele

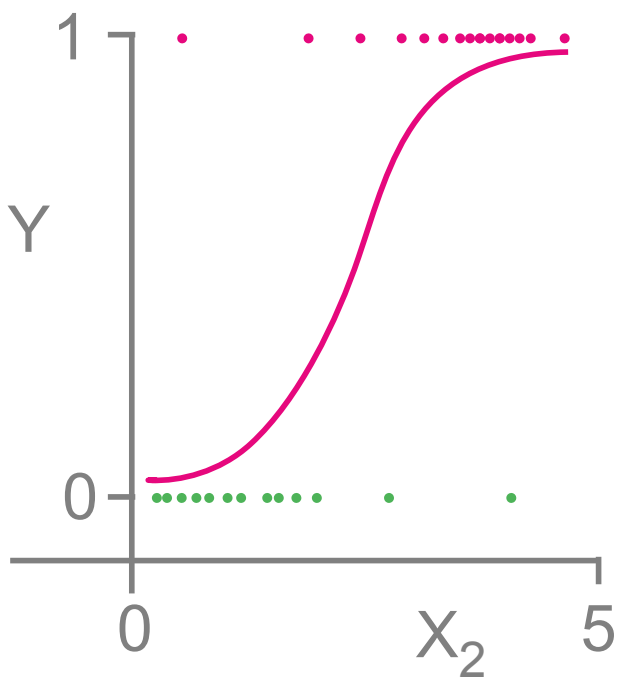
- Entscheidungen
  - etwas zu kaufen
  - wählen zu gehen
  - eine Partei wählen
- Selektion
  - Medien
  - einen Beitrag
  - Erinnerung

Lineare vs. logistische Regression

Lineare Regression



Logistische Regression



Lineare Regression vs. Logistische

3.2 Die Formel

Die Formel (können Sie gleich wieder vergessen)

$$P(y=1) = \frac{1}{1+e^{-z}} \tag{3}$$

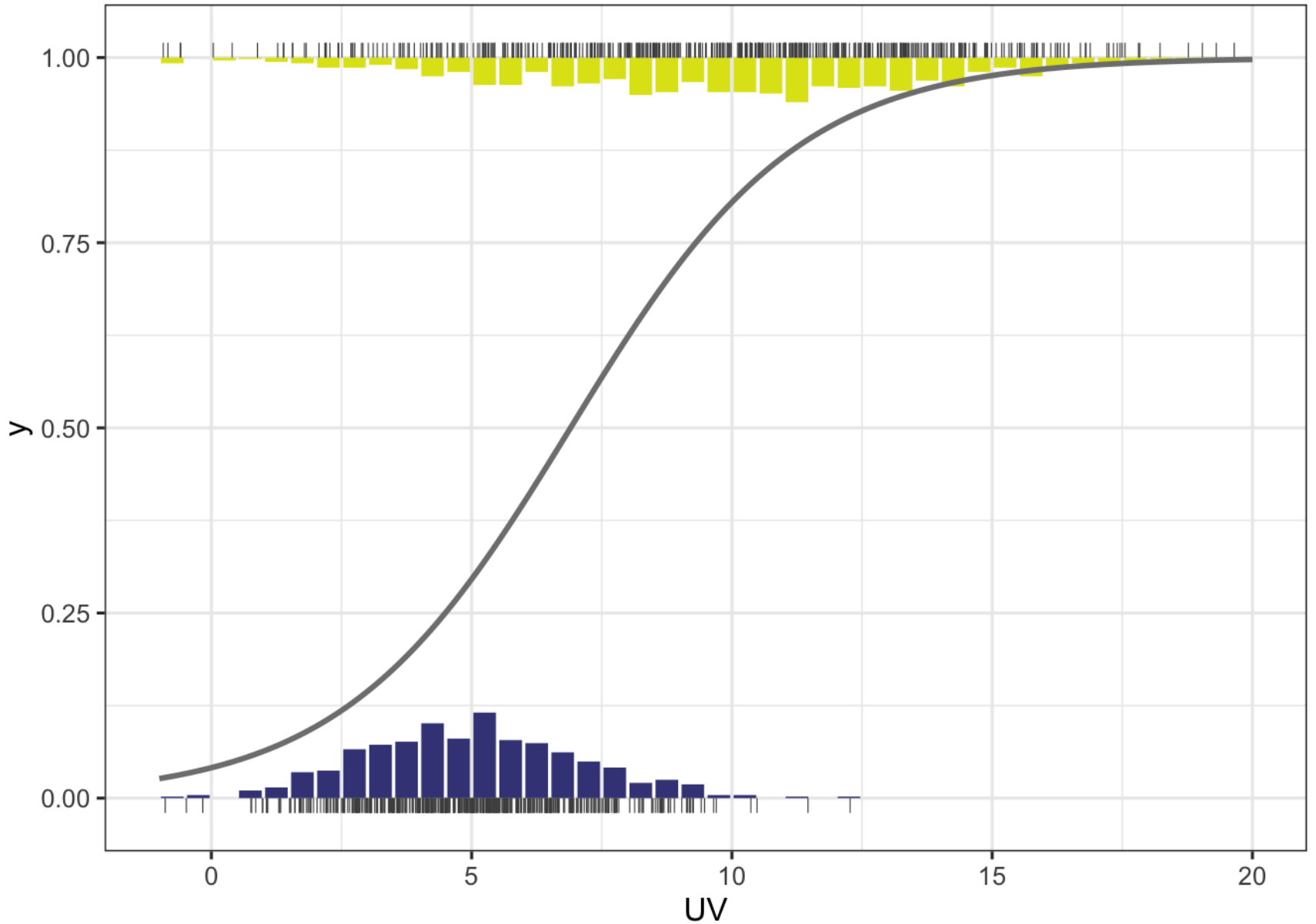
$$z = b_1 + b_2 X_{i2} + b_3 X_{i3} + e_i \tag{4}$$

$$P(Y_i) = \frac{1}{1+e^{-(b_1+b_2 X_{2i}+b_3 X_{i3}+e_i)}} \tag{5}$$

Was wenn?

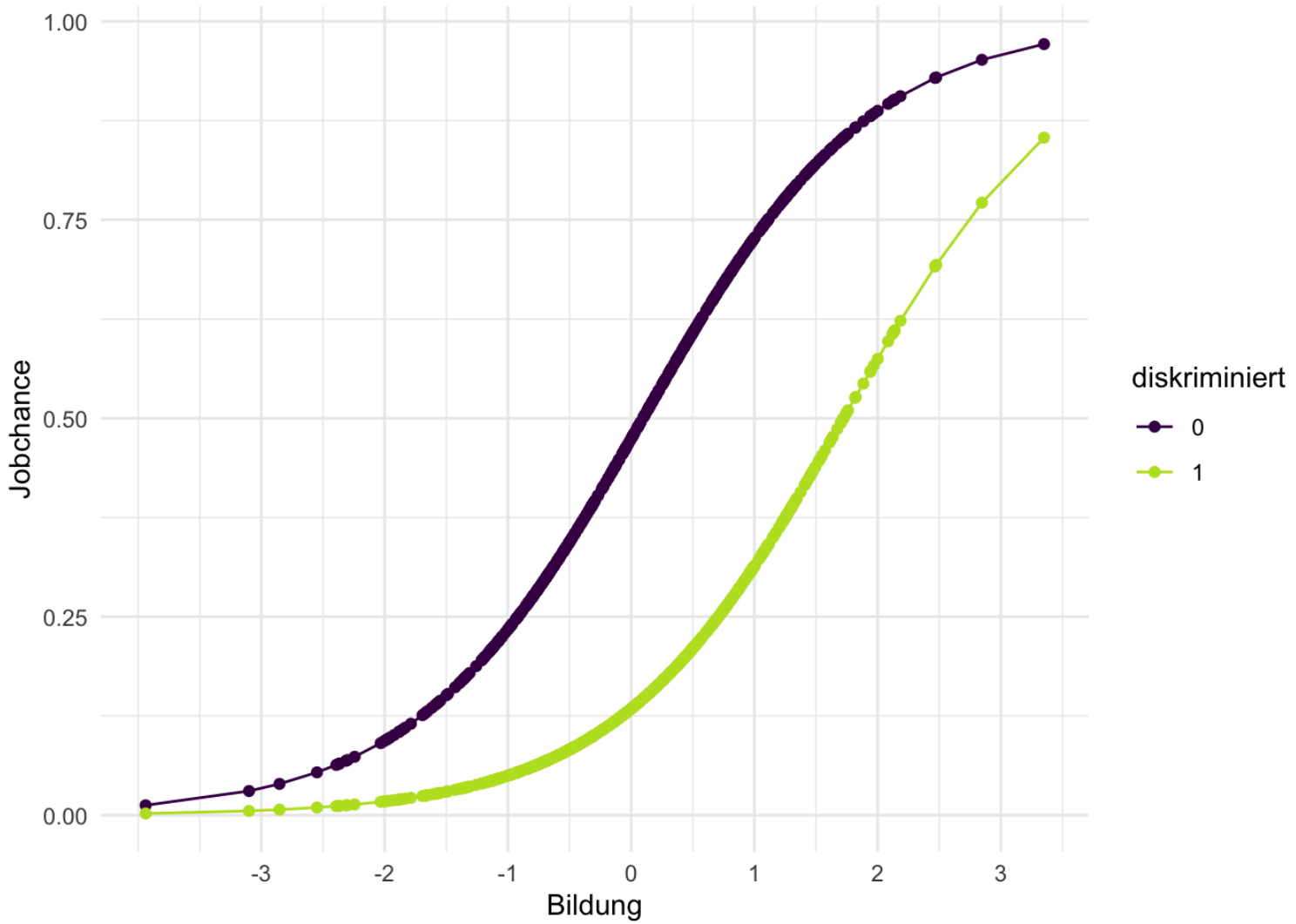
Wenn die b's gross sind und die Werte der UVs auch, dann ist $e^{-\text{Regressionsgleichung}}$ sehr klein und es bleibt 1/(1 + fast nix), also 1. Wird die Summe der Regressionsgleichung klein, wird der Nenner des Bruchs sehr gross und damit geht der Bruch gegen 0.

3.3 Grafisch



Graphisch II

$$\text{linear:} \quad \text{Jobchance} = \beta_2 \text{bild}_i + \beta_3 \text{disk}_i + \epsilon_i \quad \rightarrow \quad \text{Log Reg:} \quad \pi_i = \frac{1}{1+exp^{X\beta}}$$



Voraussetzungen

- Wie bei der linearen Regression, nehmen wir auch hier Linearität, keine Multikollinearität und Unabhängigkeit der Fehler an.
- Linearität = die UVs stehen in linearer Beziehung zum log der AV. Die AV wird also tranformiert.
- Jede mögliche Kombination möglicher Werte sollte in den Daten vorhanden sein, sonst werden die Schätzer unsicher s_b.
- Kann die AV perfekt vorhergesagt werden, nennt man das «perfekte Seperation». Die Schätzer werden unsicher s_b.

3.4 Interpretation der binär logistischen Regression

- Die B's haben die gleiche Bedeutung wie bei linearer Regression, beziehen sich aber auf die logarithmierte AV. Das Vorzeichen ist interpretierbar (positiv/negativ, wenn signifikant).
- Die Interpretation der Zusammenhänge läuft am besten über die Exp(B) = «Odds Ratio» (OR). OR geben an, wie stark sich die (Wett)Quote für AV = 1 ändert, wenn die die UV um eine Einheit grösser wird.
- Ist OR > als 1, steigt die Quote. Ist OR < 1, sinkt sie. Das CI der OR schliesst 1 ein, wenn der Effekt nicht signifikant ist.
- Die H0 der OR geht von 1 aus. Die Werte gehen von $\frac{1}{\infty}$ bis ∞.

3.5 Odds und Odds-Ratio

Odds

$$Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y_{\text{trifft ein}})}{1-P(Y_{\text{trifft ein}})}$$

Odds Ratio

$$OR = Exp(B) = e^{\beta} = \frac{\text{Odds nach dem Ansteig von x um eine Einheit}}{\text{Odds vor dem Anstieg von x um eine Einheit}} = \frac{Odds_{nach}}{Odds_{vor}}$$

$$Odds_{nach} = Exp(B) \cdot Odds_{vor}$$

B (Regressionskoeffizient)	**Exp(B)** (Odds Ratio)		**P(y=1)**
B > 0	$e^B > 1$	nimmt um den Faktor Exp(B) zu	Zunahme
B = 0	$e^B = 1$	bleibt gleich	bleibt gleich
B < 0	$e^B < 1$	sinkt um den Faktor Exp(B)	Abnahme

RAQ erklärt Interesse an Maschine Learning

LogReg: Aversion gegen Computer erklärt Interesse an ML

Characteristic	OR[1]	95% CI[1]	p-value	VIF[1]
ML1	0.74	0.53, 1.02	0.065	1.0
ML2	0.54	0.34, 0.83	0.007	1.7
ML3	1.74	1.12, 2.81	0.018	1.7

[1] OR = Odds Ratio, CI = Confidence Interval, VIF = Variance Inflation Factor

Null deviance = 230; Null df = 165; Log-likelihood = -109; AIC = 226; BIC = 238; Deviance = 218; Residual df = 162; No. Obs. = 166

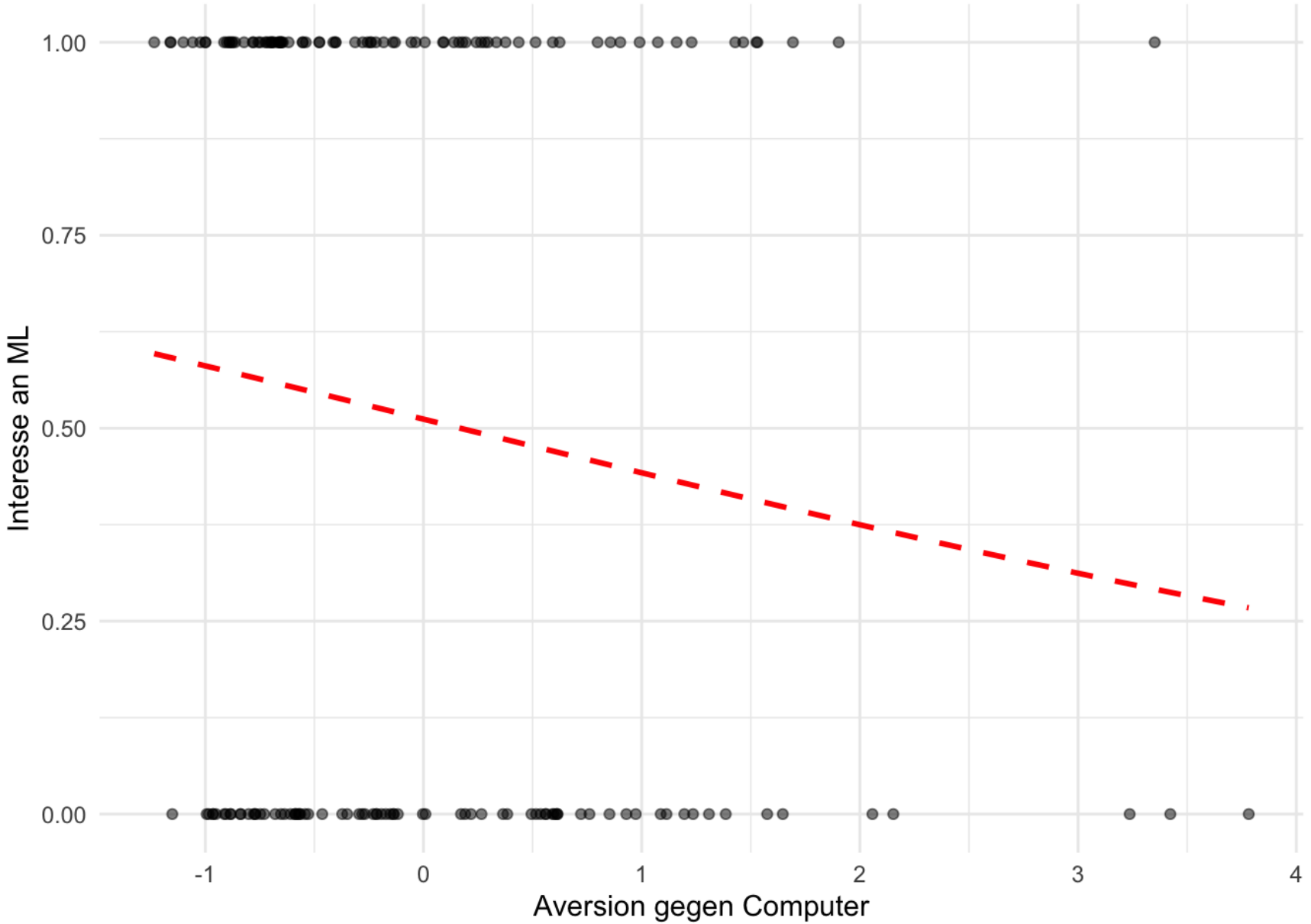
LM: Aversion gegen Computer erklärt Interesse an ML

Characteristic	Beta	95% CI[1]	p-value	VIF[1]
ML1	-0.07	-0.15, 0.00	0.066	1.0
ML2	-0.14	-0.24, -0.04	0.005	1.6
ML3	0.13	0.02, 0.23	0.016	1.6

[1] CI = Confidence Interval, VIF = Variance Inflation Factor

R^2 = 0.071; Adjusted R^2 = 0.054; Sigma = 0.488; Statistic = 4.13; p-value = 0.007; df = 3; Log-likelihood = -114; AIC = 239; BIC = 254; Deviance = 38.5; Residual df = 162; No. Obs. = 166

RAQ-ML-«S-Kurve» sieht gerade aus



4 Multinominale Regression

Multinomial bedeutet, dass es eine kategoriale (multinomiale) AV gibt. Im Prinzip werden mehrere binäre logistische Regressionen durchgeführt und zusätzlich ausgegeben, wie gut das Gesamtmodel ist. Es kommen die Fit-Kennungen AIC und BIC dazu.

Wir schauen uns nächste Woche Outputs an.

Take Home – Ausblick – Vokabeln

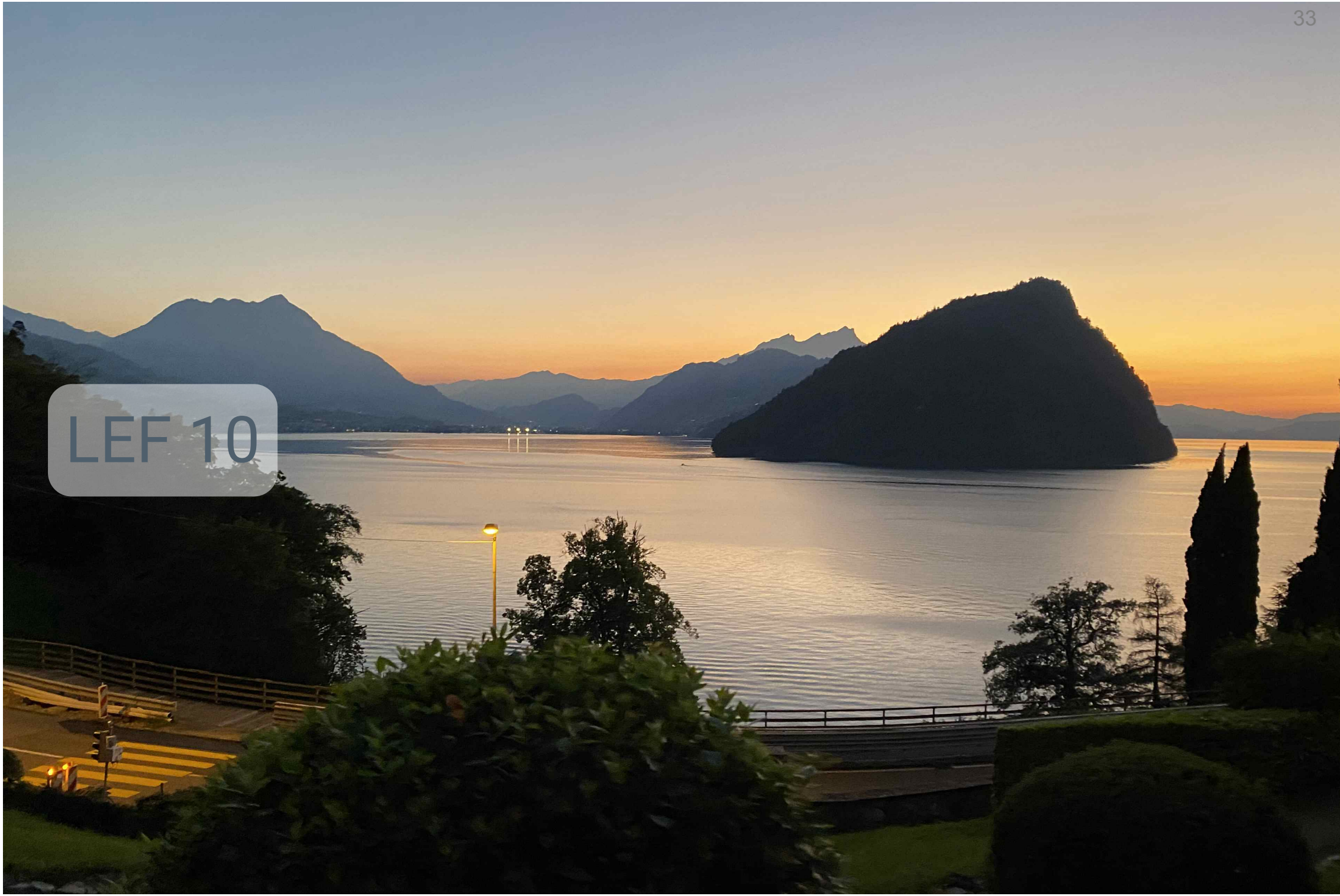
Take Home

Logistische Regression

- Wenn die AV eine Dummy ist, dann werden eher Logistische Regressionen gerechnet.
- Die Hauptkennwerte der LR sind die Exp(B) oder auch Odds Ratio (OR) genannt.
- Sie wissen, wie man die OR interpretiert
- Sie wissen, wie Machine Learning funktioniert
- Sie kennen den Unterschied zwischen «unsupervised» und «supervised» ML

Ausblick

1. Übung 4 zur Logistischen Regression und zu ML. Die Uebung_04.qmd können Sie hier herunterladen herunterladen oder Sie finden Sie auf OLAT im Materialordner.
2. Und lesen Sie den Text von Zerback Wirz! !


LEF 10

Essayfragen 10

E10.1 Was bedeuten die Exp(B) in der logistischen Regression?

E10.2 Wenn eine AV aus einer kategorialen Variable besteht mit drei Ausprägungen. Was muss man dann als Analyse damit machen?

E10.3 Die Voraussetzungen für eine normale Regression gelten auch bei der logistischen. Nennen Sie eine darüber hinausgehende Voraussetzung.

E10.4 Was bedeutet es, wenn ein Exp(B) einer logistischen Regression fast 0 ist?

E10.5 Wenn in einer Analyse die Wahlteilnahme die AV und die Frage, ob jemand Nachrichten konsumiert die UV ist, wie interpretieren Sie dann ein zugehöriges EXP(B) von 2?

E10.6 Schreiben Sie als Formel oder frei in Ihren Worten auf, was die ODDS-Ratio bedeutet.

E10.7 Was verbirgt sich hinter der Abkürzung «ML»?

E10.8 Wenn man in einem ML für das Trainingsmaterial nur einen Teil (sagen wir 70%) der gesammelten Daten nimmt, sind dann die b's nicht verzerrt? Begründen Sie Ihre Antwort. Wenn sie es nicht sind, warum nimmt man dann nicht immer nur 70% der Daten? Das wäre doch billiger?

MC-Fragen 10

MC 10.1.

MC 10.1: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Die «Odds Ratio» (OR) und Exp(B) sind dasselbe.
○	○	Logistische Regressionen werden gerechnet, wenn eine UV eine Dummy ist.
○	○	Die OR geht von -1 bis + 1.
○	○	Ein Exp(B) von 1 ist ein perfekter Zusammenhang.

Punkte: 0

MC 10.2.

MC 10.2: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Eine OR < 0 zeigt an, dass die S-Kurve von oben nach unten verläuft.
○	○	Sind UV und AV Dummys, zeigt ein EXP(B) > 1 das Vielfache an, um dass die Wahrscheinlichkeit der 1 in der AV grösser ist, wenn die UV 1 ist.
○	○	Bei perfekter Separation kann die Logistische Regression nicht gerechnet werden
○	○	Ist bei einer logistischen Regression das B signifikant, ist immer auch Exp(B) signifikant.

Punkte: 0

MC 10.3.

MC 10.3: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Soll eine kategoriale Variable mit mehreren Ausprägungen vorhergesagt werden, braucht es dafür Multinominale Modelle.
○	○	Eine Multinominale Regression bedeutet im Grunde, dass für jede Ausprägung der AV eine logistische Regression gerechnet wird.
○	○	Multinominale Regressionen werden gebraucht, wenn es nominale UV gibt.
○	○	Multinominale Regressionen werden zu den GLM gezählt.

Punkte: 0

MC 10.4.

MC 10.4: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Während bei einer linearen Regression ein R2 von .5 gross wäre, ist das für ein ML deutlich zu klein.
○	○	Beim Machine Learning mit Regressionen müssen die Voraussetzungen für OLS nicht geprüft werden.
○	○	Die PCA wird im ML häufig eingesetzt, um Vorhersagen zu treffen.
○	○	Beim ML gibt es viele andere Schätzmethoden als OLS.

Punkte: 0

MC 10.5.

MC 10.5: Sind folgende Aussagen richtig oder falsch?

richtig	falsch	Aussagen
○	○	Machine Learning wird mit ML abgekürzt.
○	○	Es gibt supervised und unsupervised ML.
○	○	Bei supervised ML wird immer mit Daten trainiert und dann an anderen Daten getestet.
○	○	ML kann mit logistischen Regressionen gemacht werden, aber nicht mit OLS Regressionen.

Punkte: 0

MC 10.6.

MC 10.6: Mal was zu R?

richtig	falsch	Aussagen
○	○	Der Schlüsselbefehl für ein lineares Modell in R lautet: lm()
○	○	Tidyverse ist ein Befehl, um R kleiner zu machen.
○	○	R-Studio ist ein Programm, mit dem R gesteuert werden kann.
○	○	Die Benennung von Kennwerten in R-Outputs ist über alle Pakete hinweg gleich.

Punkte: 0

Insgesamt 0 von 12 Punkten, was 0% und etwa einer 1 entspricht.

Vokabeln

Search:

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
All	All	All	All	All	All
89	8	ML, Logit-Modelle, Multinominal	Datensplit	data splitting	Aufteilung der Daten in zwei Datensätze (zB Trainings- und Test-Daten).
90	8	ML, Logit-Modelle, Multinominal	Exponentielle Beta	exponential beta	In der Logistischen Regression die Odds Ratio, also die Quote für das Ergebnis 1.
91	8	ML, Logit-Modelle, Multinominal	Feature Engineering	Feature Engineering	Die Suche und transformation von Prädiktoren für eine AV im ML.
92	8	ML, Logit-Modelle, Multinominal	Generalisiertes Lineares Modell	Generalized Linear Model (GLM)	Beim GLM werden nichtlineare Beziehungen modelliert (auch Logit)

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
					Heteroskedastizität durch Gewichtungen ausgeglichen.
93	8	ML, Logit-Modelle, Multinominal	Logistische Regression	logistic regression	Regression mit einer Dummy als AV.
94	8	ML, Logit-Modelle, Multinominal	Machine Learning (ML)	machine learning (ML)	Machine Learning ist ein künstlicher Prozess bei dem Computer anhand von vorgegebenen Parametern und Datenmaterial Wissen generieren, indem sie Algorithmen auf grosse Datenmengen anwenden.
95	8	ML, Logit-Modelle, Multinominal	Machine Learning, reinforcement	machine learning, reinforcement	Vom System werden Möglichkeiten probiert, wobei ein Ziel vorgegeben ist und der jeweilige Zustand positiv oder negativ sanktioniert wird.

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
96	8	ML, Logit-Modelle, Multinominal	Machine Learning, supervised	machine learning, supervised	Beim Supervised Machine Learning werden wird das System trainiert, indem Untersuchungsgegenstände und Ergebnisse als Trainingsmaterial vorgegen werden.
97	8	ML, Logit-Modelle, Multinominal	Machine Learning, unsupervised	machine learning, unsupervised	Beim Unsupervised Machine Learning werden keine Vorgaben gemacht. Das System sucht (und findet) selbständig Muster und Strukturen in Daten.
98	8	ML, Logit-Modelle, Multinominal	Multinominale Regression	multinominal logistic Regression	Wird angewendet, wenn eine kategoriale Variable als AV genommen wird. Je Ausprägung der AV wird im Grunde eine logistische Regression gerechnet.

Nr	Sitzung	Inhalt	Deutsch	Englisch	Erläuterung
99	8	ML, Logit-Modelle, Multinominal	Quote	Odds Ratio (OR)	Die Quote dafür, dass die 1 als Ergebnis steht und nicht die 0. Ist die OR > 1, steigt die Wahrscheinlichkeit für 1. Wenn OR < 1, dann sinkt die Wahrscheinlichkeit.
101	8	ML, Logit-Modelle, Multinominal	Testdaten	test data	Daten an denen ein Modell getest wird.
102	8	ML, Logit-Modelle, Multinominal	Trainingsdaten	train data	Daten an denen ein Modell trainiert wird.
103	8	ML, Logit-Modelle, Multinominal	Unterspezifikation	under fitting	Bedeutet, dass wichtige Prädiktoren vorhanden wären, aber nicht spezifiziert wurden.
100	8	ML, Logit-Modelle, Multinominal	Überspezifikation	over fitting	Beudetet, dass zu viel angepasst wurde und Prognosen mit irrelavanten