Anwendungsorientierte Analyseverfahren

Logistische Regression

Prof. Dr. Michael Scharkow

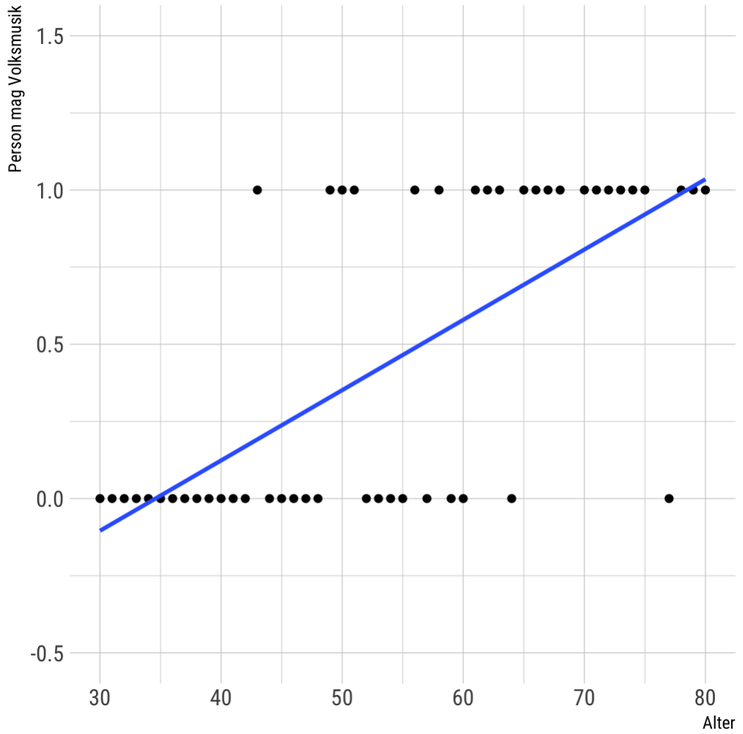
Sommersemester 2024

Kategorielle Outcomes - χ^2 Test

- oft sind wir an Zusammenhängen von kategoriellen Variablen interessiert
- Klassische Analysestrategie: Kreuztabelle und χ^2 Test auf Unabhängigkeit
- Vorteil: einfache Darstellung (Tabelle mit Spaltenprozenten, χ^2 Wert, p-Wert, Kontingenzkoeffizient)
- Nachteil: der χ^2 Test ist ein Globaltest, d.h. wir testen nur, ob es irgendwo signifikante Abweichungen der beobachteten von den erwarteten Häufigkeiten gibt, aber nicht wie und wo
- weiterer Nachteil: nur kategorielle Prädiktoren, nur bivariate Zusammenhänge
- Alternative: GLM mit logistischer Link-Funktion

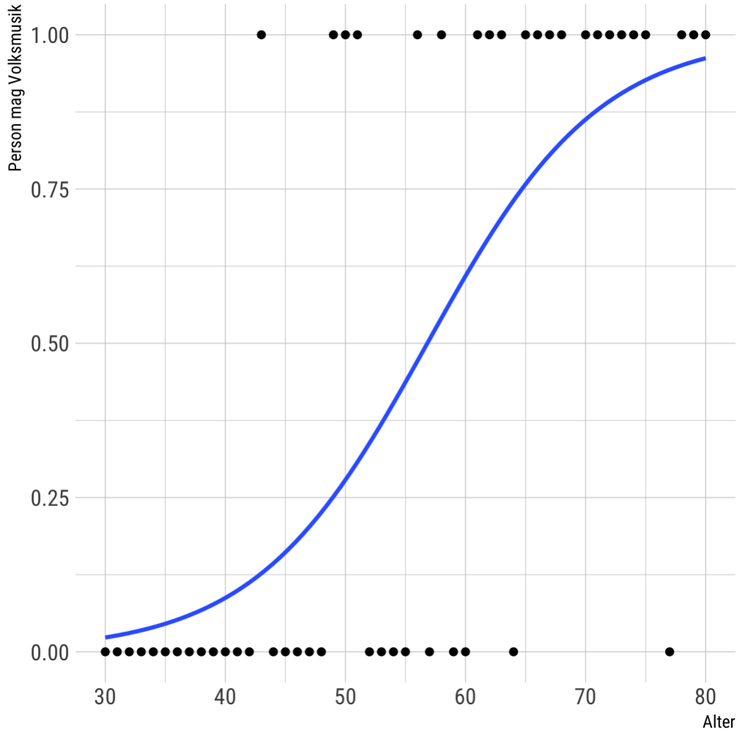
LINEARE REGRESSION BEI DICHOTOMER AV

- viele vorausgesagte Werte von Y kann es in den Daten nicht geben, da nur die Werte 0 und 1 empirisch vorkommen.
- linearer Zusammenhang ist nicht gegeben, d.h. wir verstoßen gegen die Annahmen der OLS-Regression
- trotzdem häufiger Einsatz als Linear Probability Model (LPM) mit einfacher Interpretation



(INVERSE) LOGIT-FUNKTION

- Logit-Funktion: $P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
- Die Logit-Funktion transformiert alle Werte in einen Bereich von 0-1.
- Die Werte von 0-1 können als Wahrscheinlichkeiten interpretiert werden].



LINEARE VS. LOGISTISCHE REGRESSION

$$\text{Logit}(Y) = b_0 + b_1 X_i + \epsilon_i$$

- gleiche Annahmen zu Unabhängigkeit und Multikollinearität
- gleiche Logik der Modellspezifikation mit Nullmodell und Prädiktorvariablen
- gleiche Logik der statistischen Inferenz, d.h. Standardfehler und Konfidenzintervalle
- unterschiedliche Interpretation der Koeffizienten (sowohl unstandardisiert als auch standardisiert)
- unterschiedliche Maße der Modellgüte (R^2, etc.)

INTERPRETATION LOGISTISCHER MODELLE

- unstandardisierte Koeffizienten sind bei logistischer Regression schwer interpretierbar
- Interpretation als (Änderung von) Wahrscheinlichkeiten ist **falsch**, u.a. weil diese nicht konstant sind
- 3 Möglichkeiten, logistische Modelle zu interpretieren:
  - Divide-by-4-Regel mit unstandardisierten Koeffizienten
  - Average Marginal Effects
  - Odds-Ratios

UNSTANDARDISIERTE KOEFFIZIENTEN

- die Interpretation von unstandardisierten Koeffizienten im logistischen Modell hängt (auch) am Intercept bzw. an den anderen Kovariaten
- der Intercept lässt sich durch die sog. Inverse Logit Funktion (in R `plogis()`) in eine Baseline-Wahrscheinlichkeit umrechnen
- die Veränderung der Wahrscheinlichkeiten bei einer Änderung von X (Slope) ist nicht in allen Fällen gleich groß
- Gelman & Hill (2007) empfehlen für eine schnelle Interpretation unstandardisierte Koeffizienten die Divide-by-4-Regel
- $B/4$ ist eine *Obergrenze* für die Änderung in der Wahrscheinlichkeit P(Y=1), wenn X sich um eine Einheit ändert
- Abweichung ist am Rande der Verteilung von Y größer als in der Mitte

AVERAGE MARGINAL EFFECTS

- für jeden Wert von X lässt sich der Anstieg in der Logit-Funktion von Y berechnen.
- wenn wir dies für alle Werte in der Stichprobe tun, und davon den Mittelwert errechnen, erhalten wir den AME
- dieser lässt sich als durchschnittliche Veränderung in der Wahrscheinlichkeit P(Y=1) über alle Fälle interpretieren
- alternativ können wir auch wieder typische Fälle auswählen und dafür die vorgesagten Werten von Y schätzen
- dies bietet sich vor allem bei Modellen mit kategoriellen Prädiktorvariablen an (vgl. ANOVA)

ODDS RATIOS - EXP(B)

- die meisten logistischen Regressionen berichten statt B den Wert Exp(B), der auch als Odds Ratio (OR) bezeichnen wird
- Odds = $P(x)/1 - P(x)$ kann man im Deutschen als Chance oder Risiko übersetzen, nicht jedoch als Wahrscheinlichkeit!
- eine OR=1 bedeutet kein Effekt, OR > 1 bedeutet eine *erhöhte* Chance bzw. Risiko, eine OR < 1 eine *niedrigere* Chance/Risiko
- Exp(B) = .5 heißt: mit jedem Skalenpunkt mehr X besteht ein halb so großes Risiko
- Exp(B) = 2 heißt: mit jedem Skalenpunkt mehr X verdoppelt sich das Risiko

MODELLGÜTE: PSEUDO R^2

- weit verbreitetes Maß für die Güte einer logistischen Regression ist das Pseudo R^2 von McFadden, Nagelkerke oder Cox & Snell.
- diese setzen die Log-Likelihood des Nullmodells mit dem gefitteten Modell in Beziehung, beschreiben also wieviel besser das Modell gegenüber dem Nullmodell ist
- die Interpretation ist vergleichbar mit dem klassischen R^2, wobei nicht immer klar ist, inwiefern der Wertebereich wirklich von 0-1 geht

Beispielstudie Festl et al. (2013)

PEER INFLUENCE, INTERNET USE AND CYBERBULLYING: A COMPARISON OF DIFFERENT CONTEXT EFFECTS AMONG GERMAN ADOLESCENTS

Ruth Festl, Michael Scharkow and **Thorsten Quandt**

The influence of social reference groups such as family members, classmates and friends on adolescents' attitudes and behavior has been acknowledged in research for many decades. With the increasing use of online media, cyberbullying has become a major issue in adolescence research. In this paper, we compare various forms of peer influence on cyberbullying behavior among high school students in Germany. Specifically, the impact of close friends and more distant peers in the school class on perpetrator and victim roles is compared. The results indicate that the class context is highly relevant for cyberbullying. For both processes—perpetration and victimization—the number of cyberbullies within a school class plays an important role in predicting individual behavior. Looking at individual risk factors, the results show that cyberbullying is strongly related to the use of social networking sites, and the risk of victimization increases with the time spent online.

DATEN

cbbully_gen	cbvictim_gen	age	gender	internetuse	class
0	0	17	female	3.0	G11c
0	1	18	female	8.0	G12
0	0	18	male	3.0	G12
0	0	15	male	2.5	G9b
1	0	19	male	8.5	G11c

Variable	n_Obs	Mean	SD	Median	MAD	Min	Max
cbvictim_gen	275	0.11	0.31	0	0	0	1

NULLMODELL

Parameter	Coefficient	95% CI	z	p	Std. Coef.	Fit
(Intercept)	-2.14	(-2.52, -1.75)	-10.89	< .001	-2.14	
AICc						187.31
Tjur's R2						0.00
Sigma						1.00
Log_loss						0.34

```
plogis(-2.14)
```

```
[1] 0.1052694
```

KATEGORIELLER PRÄDIKTOR

Parameter	Coefficient	95% CI	z	p	Std. Coef.	Fit
(Intercept)	-2.48	(-3.13, -1.94)	-8.27	< .001	-2.48	
gender (female)	0.69	(-0.08, 1.50)	1.74	0.082	0.69	
AICc						186.26
Tjur’s R2						0.01
Sigma						1.00
Log_loss						0.33

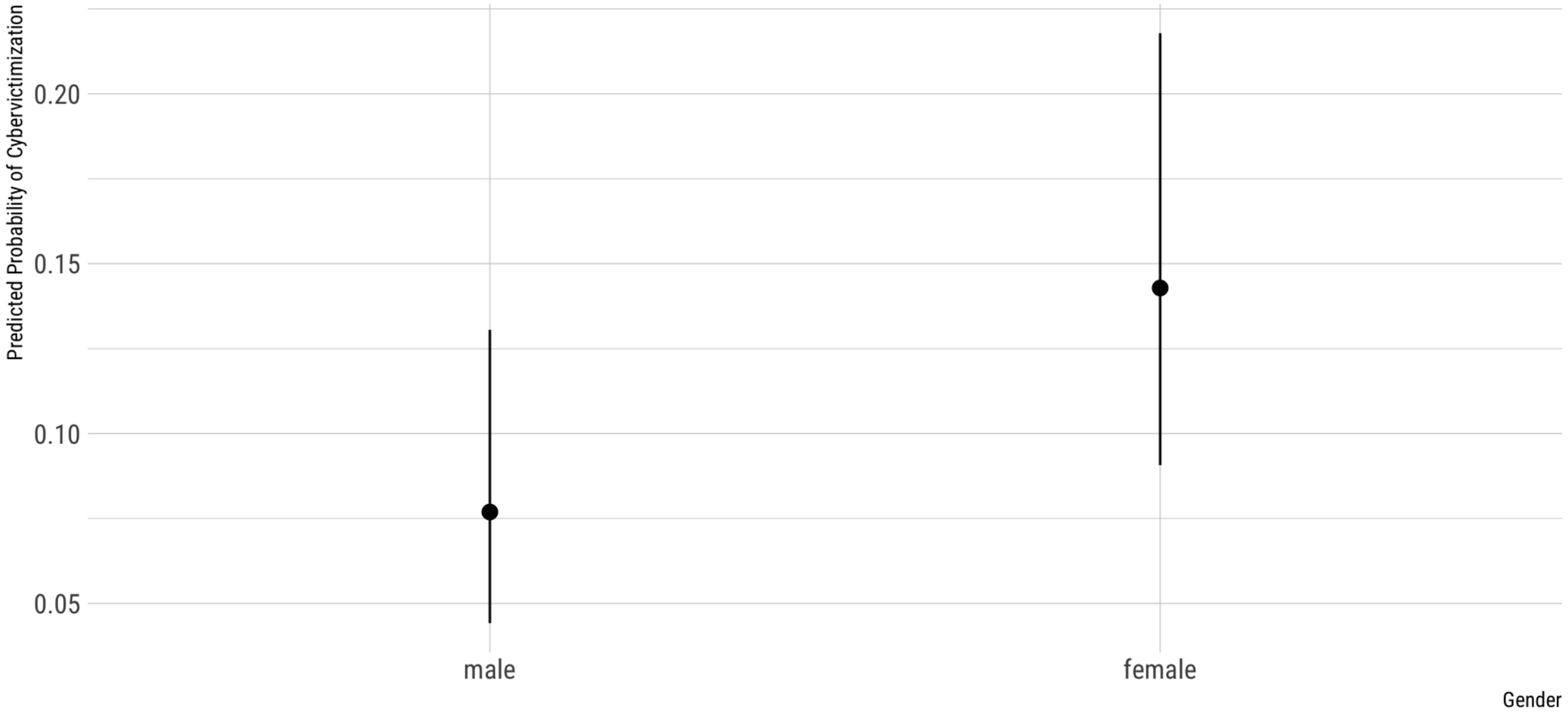
ODDS-RATIOS (EXP(B))

	Parameter	Coefficient	CI	CI_low	CI_high	z	p	Fit
1	(Intercept)	0.08	0.95	0.04	0.14	-8.27	0.00	
2	gender [female]	2.00	0.95	0.92	4.46	1.74	0.08	
3								
4	AIC							186.22
5	AICc							186.26
6	BIC							193.45
7	Tjur’s R2							0.01
9	Sigma							1.00
10	Log_loss							0.33

AVERAGE MARGINAL EFFECTS

term	estimate	std.error	statistic	p.value	conf.low	conf.high
gender	0.07	0.04	1.71	0.09	-0.01	0.14

VORHERGESAGTE WAHRSCHEINLICHKEITEN



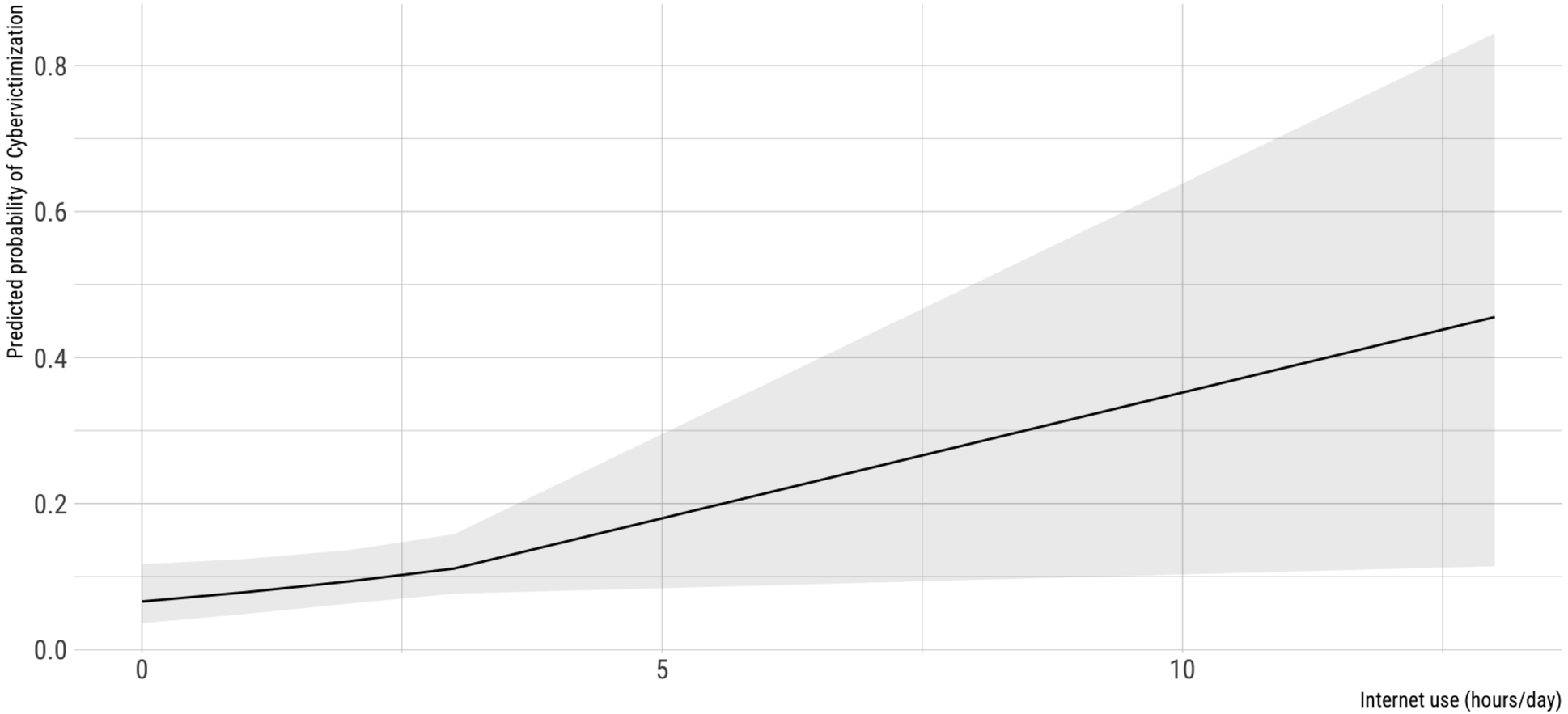
MULTIPLE LOGISTISCHE REGRESSION

Parameter	Coefficient	95% CI	z	p	Std. Coef.	Fit
(Intercept)	-2.98	(-3.82, -2.23)	-7.41	< .001	-2.56	
gender (female)	0.76	(-0.02, 1.58)	1.88	0.060	0.76	
internetuse	0.19	(0.00, 0.37)	2.10	0.036	0.34	
AICc						184.36
Tjur’s R2						0.03
Sigma						1.00
Log_loss						0.32

AVERAGE MARGINAL EFFECTS

term	estimate	std.error	statistic	p.value	conf.low	conf.high
gender	0.07	0.04	1.85	0.06	0	0.15
internetuse	0.02	0.01	1.98	0.05	0	0.04

MODELLVORHERSAGEN



FAZIT

- in vielen Fällen ist die logistische Regression eine sinnvolle Alternative zu Kreuztabellen
- die Modellierung ist praktisch identisch mit linearen Regressionsmodellen
- die Interpretation der Koeffizienten ist komplexer, aber AME und Modellvorhersagen helfen dabei
- das Linear Probability Model führt oft zu ähnlichen Ergebnissen (vor allem wenn Y nicht allzu selten oder häufig 1 ist)

LITERATUR

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

Festl, R., Scharkow, M., & Quandt, T. (2013). Peer influence, internet use and cyberbullying: A comparison of different context effects among German adolescents. Journal of Children and Media, 7(4), 446-462.